REPORT DOCUMENTATION PAGE

Form Approved
OMB NO. 0704-0188

Public Reporting burden for this collection of information is estimated to average 1 hour per response, including the time for reviewing instructions, searching existing data sources, gathering and maintaining the data needed, and completing and reviewing the collection of information. Send comment regarding this burden estimates or any other aspect of this collection of information, including suggestions for reducing this burden, to Washington Headquarters Services, Directorate for information Operations and Reports, 1215 Jefferson Davis Highway, Suite 1204, Arlington, VA 22202-4302, and to the Office of Management and Budget, Paperwork Reduction Project (0704-0188,) Washington, DC 20503.

1. AGENCY USE ONLY (Leave Blank)	2. REPORT DATE	3. REPORT TYPE AND DATES COVERED
	12/24/2009	Final 03/02/2008 - 11/30/2009

4. TITLE AND SUBTITLE

Analysis and Design of Manyeore Processor-to-DRAM Opto-Eleetrieal Networks with Integrated Silieon Photonies

5. FUNDING NUMBERS

W911NF-08-1-0134 and W911NF-08-1-0139

6. AUTHOR(S)

Vladimir Stojanovie and Krste Asanovie

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)

MIT, 77 Mass Ave, Cambridge, MA 02139
ICSI, 1947 Center Street. Suite 600, Berkeley, CA 94704

8. PERFORMING ORGANIZATION REPORT NUMBER

9. SPONSORING / MONITORING AGENCY NAME(S) AND ADDRESS(ES)

U. S. Army Research Office
P.O. Box 12211
Research Triangle Park, NC 27709-2211

10. SPONSORING / MONITORING AGENCY REPORT NUMBER

11. SUPPLEMENTARY NOTES

The views, opinions and/or findings contained in this report are those of the author(s) and should not be construed as an official Department of the Army position, policy or deeision, unless so designated by other doeumentation.

12 a. DISTRIBUTION / AVAILABILITY STATEMENT

Approved for public release; distribution unlimited.

12 b. DISTRIBUTION CODE

13. ABSTRACT (Maximum 200 words)

In this report we summarize the results of the investigation to improve the performanee of manyeore processors with silieon photonie networks for eore-to-eore and DRAM communieation. Our findings indieate that in on-ehip eore-to-eore networks, photonie Clos networks improve the energy-efficieney by 2-3x over other types of electrieal networks while providing uniform bandwidth and lateney over a range of different applieations, at the fraetion of the eosts and with orders of magnitude smaller demands on photonie deviee eount, power and footprint than global photonie erossbar networks.
We also propose a way to utilize the photonie intereonneets to effieiently eonneet multi-soeketed ehips into a flat, shared memory network, enabling a higher degree of die disintegration for improvements in yield, paekaging, thermal and power delivery.

20091231177

14. SUBJECT TERMS			15. NUMBER OF PAGES
			16. PRICE CODE
17. SECURITY CLASSIFICATION OR REPORT	18. SECURITY CLASSIFICATION ON THIS PAGE	19. SECURITY CLASSIFICATION OF ABSTRACT	20. LIMITATION OF ABSTRACT
UNCLASSIFIED	UNCLASSIFIED	UNCLASSIFIED	UL

NSN 7540-01-280-5500

Standard Form 298 (Rev.2-89)
Prescribed by ANSI Std. 239-18
298-102

SF 298 Continuation Sheet

1) List of papers submitted or published under ARO sponsorship during the grant period from 03/02/2008-11/30/09.

a) Manuscripts submitted but not published

None.

b) Papers published in peer-reviewed journals

None.

c) Papers published in non-peer reviewed journals or conference proceedings

Joshi, A., C. Batten, Y-J. Kwon, S. Beamer, K. Asanović, and V. Stojanović, "Silicon-Photonic Clos Networks for Global On-Chip Communication," 3rd ACM/IEEE International Symposium on Networks-on-Chip, San Diego, CA, pp. 124-133, May 2009.

Stojanović, V., A. Joshi, C. Batten, Y-J. Kwon, K. Asanović, "Manycore Processor Networks with Monolithic Integrated CMOS Photonics," *Optical Society of America - CLEO/QELS Conference* , Baltimore, MD, 2 pages, June 2009.

Beamer, S., K. Asanović, C. Batten, A. Joshi, and V. Stojanović, "Designing multi-socket systems using silicon photonics," in *Proceedings of the 23rd International Conference on Supercomputing*, Yorktown Heights, NY, pp. 521-522, June 2009.

d) Papers presented at meetings, but not published in conference proceedings

None.

2) Demographic Data for this Reporting Period:

a) Number of Manuscripts submitted during this reporting period: 3

b) Number of Peer Reviewed Papers submitted during this reporting period: 3

c) Number of Non-Peer Reviewed Papers submitted during this reporting period: None.

d) Number of Presented but not Published Papers submitted during this reporting period: None.

3) Demographic Data for the life of this agreement:

a) Number of Scientists Supported by this agreement: None.

b) Number of Inventions resulting from this agreement: None

c) Number of PhD(s) awarded as a result of this agreement: None

d) Number of Bachelor Degrees awarded as a result of this agreement: None

e) Number of Patents Submitted as a result of this agreement: None.

f) Number of Patents Awarded as a result of this agreement: None.

g) Number of Grad Students supported by this agreement:

h) Number of Grad Students supported by this agreement: 4

i) Number of FTE Grad Students supported by this agreement: None.

j) Number of Post Doctorates supported by this agreement: 1

k) Number of FTE Post Doctorates supported by this agreement: None.

l) Number of Other Staff supported by this agreement: None.

m) Number of Undergrads supported by this agreement: None.

n) Number of Master Degrees awarded as a result of this agreement: None.

4) Student Metrics for graduating undergraduates funded by this agreement: N/A.

5) "Report of inventions" (by title only)

6) "Scientific progress and accomplishments": N/A (see technical write-up)

7) "Technology transfer": N/A.

Final Technical Report

Report Title:	Analysis and Design of Manycore Processor-to-DRAM Opto-Electrical Networks with Integrated Silicon Photonics
Technical Point of Contact:	Professor Vladimir Stojanovic Massachusetts Institute of Technology 77 Massachusetts Ave., Room 36-260 Cambridge, MA 02139 Phone: 617-324-4913, Fax: 617-324-0862 Electronic Email: vlada@mit.edu
Administrative POC:	Ms. Mary A. McGonagle Massachusetts Institute of Technology Office of Sponsored Programs 77 Massachusetts Ave., Room E19-750 Cambridge, MA 02139 Phone: 617-258-8017, Fax: 617-253-4734 Electronic Email: mam@mit.edu

Table of Contents

I. Executive Summary

In this project we have investigated the architecture and design of manycore processor-to-DRAM networks using integrated silicon photonics. We have focused primarily on two types of networks, on-die core-to-core network, and core-to-memory controller network that possibly connects several processor sockets into a seamless, flat shared-memory systems.

In the context of core-to-core networks, we have explored the constraints on photonic technology imposed by the implementation of non-blocking networks such as crossbars and Clos. We developed a comprehensive modeling framework for estimation of optical and electrical power requirements for various physical network topologies. We have shown that in an example 64-tile system photonic Clos network consumes significantly less optical power, thermal tuning power and area, compared to global photonic crossbars, over a range of photonic device parameters. The results from our network simulation framework indicate that compared to various other electrical on-chip networks, photonic Clos networks can provide more uniform latency and throughput across a range of traffic patterns while consuming less power. These properties will help simplify parallel programming by allowing the programmer to ignore network topology during optimization. The first part of the report includes our publication of these findings, presented at the International Symposium for Networks on Chip in May 2009.

In the context of multi-socket core-to-memory controller networks we explored the use of silicon photonics to build relatively flat, high bandwidth memory interconnect. In this work, we present a scalable and coherent multi-socket design along with discussing the tradeoffs facing an architect when incorporating silicon photonics technology. This work also points to an important indirect impact of using efficient interconnect technology like silicon photonics – the impact on yield and size of processor chips. By using the efficient photonic interconnect, the motivation to integrate cores into large processor chips disintegrates, leaving room for die-size optimization to support yield improvements, ease of packaging, cooling and power delivery. Details of this work are provided in the second part of the technical report, presented at the International Conference on Supercomputing in June 2009.

II. Detailed Technical Information

A. Design of Non-Blocking Core-to-Core Photonic Networks

Silicon-Photonic Clos Networks for Global On-Chip Communication

Ajay Joshi[*], Christopher Batten[*], Yong-Jin Kwon[†], Scott Beamer[†], Imran Shamim[*]
Krste Asanović[†], Vladimir Stojanović[*]

[*] *Department of EECS, Massachusetts Institute of Technology, Cambridge, MA*
[†] *Department of EECS, University of California, Berkeley, CA*

Abstract

Future manycore processors will require energy-efficient, high-throughput on-chip networks. Silicon-photonics is a promising new interconnect technology which offers lower power, higher bandwidth density, and shorter latencies than electrical interconnects. In this paper we explore using photonics to implement low-diameter non-blocking crossbar and Clos networks. We use analytical modeling to show that a 64-tile photonic Clos network consumes significantly less optical power, thermal tuning power, and area compared to global photonic crossbars over a range of photonic device parameters. Compared to various electrical on-chip networks, our simulation results indicate that a photonic Clos network can provide more uniform latency and throughput across a range of traffic patterns while consuming less power. These properties will help simplify parallel programming by allowing the programmer to ignore network topology during optimization.

1. Introduction

Today's graphics, network, embedded and server processors already contain many processor cores on one chip and this number is expected to increase with future scaling. The on-chip communication network is becoming a critical component, affecting not only performance and power consumption, but also programmer productivity. From a software perspective, an ideal network would have uniformly low latency and uniformly high bandwidth. The electrical on-chip networks used in today's multicore systems (e.g., crossbars [8], meshes [3], and rings [11]) will either be difficult to scale to higher core counts with reasonable power and area overheads or introduce significant bandwidth and latency non-uniformities. In this paper we explore the use of silicon-photonic technology to build on-chip networks that scale well, and provide uniformly low latency and uniformly high bandwidth.

Various photonic materials and integration approaches have been proposed to enable efficient global on-chip communication, and several network architectures (e.g., crossbars [7, 15] and meshes [13]) have been developed bottom-up using fixed device technology parameters as drivers. In this paper, we take a top-down approach by driving the photonic device requirements based on the projected network and system needs. This allows quick design-space exploration at the network level, and provides insight into which network topologies can best harness the advantages of photonics at different stages of the technology roadmap.

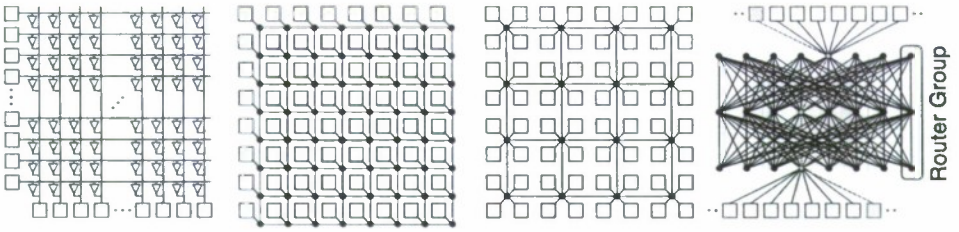
This paper begins by identifying our target system and briefly reviewing the electrical on-chip networks which will serve as a baseline for our photonic network proposals. We then use analytical models to investigate the tradeoffs between various implementations of global photonic crossbars found in the literature and our own implementations of photonic Clos networks. We also use simulations to compare the photonic Clos network to electrical mesh and Clos networks. Our results show that photonic Clos networks consume significantly less optical laser power, thermal tuning power, and area as compared to photonic crossbar networks, and offer better energy-efficiency than electrical networks while providing more uniform performance across various traffic patterns.

2. Target System

Silicon-photonic technology for on-chip communication is still in its formative stages, but with recent technology advances we project that photonics might be viable in the late 2010's. This makes the 22 nm node a reasonable target process technology for our work. By then it will be possible to integrate hundreds of cores onto a single die. To simplify design and verification complexity, these cores and/or memory will most likely be clustered into tiles which are then replicated across the chip and interconnected with a well-structured on-chip network. The exact nature of the tiles and the inter-tile communication paradigm are still active areas of research. The tiles might be homogeneous with each tile including both some number of cores and a slice of the on-chip memory, or the tiles might be heterogeneous with a mix of compute and memory tiles. The global on-chip network might be used to implement shared memory, message passing, or both. Regardless of their exact configuration, however, all future systems will require some form of on-chip network which provides low-latency and high-throughput commu-

978-1-4244-4143-3/09/$25.00 ©2009 IEEE

Authorized licensed use limited to: MIT Libraries. Downloaded on December 1, 2009 at 00:13 from IEEE Xplore. Restrictions apply.



(a) Crossbar (b) Mesh (c) CMesh (d) Clos

Figure 1: Logical View of 64 Tile Network Topologies – (a) 64x64 distributed tristate global crossbar, (b) 2D 8x8 mesh, (c) concentrated mesh (cmesh) with 4x concentration, (d) 8-ary, 3-stage Clos network with eight middle routers. In all four figures: squares = tiles, dots = routers, triangles = tristate buffers. In (b) and (c) inter-dot lines = two opposite direction channels. In (a) and (d) inter-dot lines = uni-directional channels.

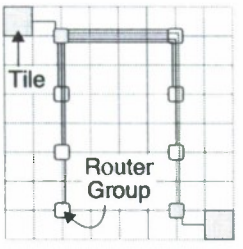


Figure 2: Clos Layout – Router group is three routers. Only a subset of the channels are shown.

nication at low energy and small area.

For this paper we assume a target system with 64 square tiles operating at 5 GHz on a 400 mm^2 chip. Figure 1 illustrates some of the topologies available for implementing on-chip networks. They range from high-radix, low-diameter crossbar networks to low-radix, high-diameter mesh networks. We examine networks sized for low (LTBw), medium (MTBw), and high (HTBw) bandwidth which correspond to ideal throughputs of 64, 128, and 256 b/cycle per tile under uniform random traffic. Although we primarily focus on a single on-chip network, our exploration approach is also applicable to future systems with multiple physical networks.

3. Electrical On-Chip Networks

In this section, we explore the qualitative trade-offs between various network architectures that use traditional electrical interconnect. This will provide an electrical baseline for comparison, and also yield insight into the best way to leverage silicon photonics.

3.1. Electrical Technology

The performance and cost of on-chip networks depend heavily on various technology parameters. For this work we use the 22 nm predictive technology models [16] and interconnect projections from [6] and the ITRS.

All of our inter-router channels are implemented in semi-global metal layers with standard repeated wires. For medium length wires (2–3 mm or approximately the width of a tile) the repeater sizing and spacing are chosen so as to minimize the energy for the target cycle-time. Longer wires are energy optimized as well as pipelined to maintain throughput. The average energy to transmit a bit transition over a distance of 2.5 mm in 200 ps is roughly 160 fJ, while the fixed link cost due to leakage and clocking is ≈20 fJ per cycle. The wire pitch is only 500 nm, which means that ten thousand wires can be supported across the bisection of our target chip even with extra space for power distribution and vias. Given the abundance of on-chip wiring resources, interconnect power dissipation will likely be a more serious constraint than bisection bandwidth for most network topologies.

We assume a relatively simple router microarchitecture which includes input queues, round-robin arbitration, a distributed tristate crossbar, and output buffers. The routers in our multihop networks have similar radices, so we fix the router latency to be two cycles. For a 5×5 router with 128 b flits of uniformly random data, we estimate the energy to be 16 pJ/flit. Notice that sending a 128 b flit across a 2.5 mm channel consumes roughly 13 pJ, which is comparable to the energy required to move this flit through a simple router. Future on-chip network designs must therefore carefully consider *both* channel and router energy, and to a lesser extent area.

3.2. Electrical On-chip Networks

Figure 1 illustrates four topologies that we will be discussing in this section and throughout the paper: global crossbars, two-dimensional meshes, concentrated meshes, and Clos networks. Table 1 shows some key parameters for these topologies assuming a MTBw system.

For systems with few tiles, a simple global crossbar is one of the most efficient network topologies and presents a simple performance model to software [8]. Such crossbars are strictly non-blocking; as long as an output is not oversubscribed every input can send messages to its desired output without contention. Small crossbars can have very low-latency and high-throughput but are difficult to scale to tens or hundreds of tiles.

Figure 1a illustrates a 64×64 crossbar network implemented with distributed tristate buses. Although such a network provides strictly non-blocking connectivity, it also requires a large number of global buses across the length of the chip. These buses are challenging to layout and must be pipelined for good throughput. Global arbitration can add significant latency and also needs to be pipelined. These global control and data wires result in significant power consumption even for communication

Authorized licensed use limited to: MIT Libraries. Downloaded on December 1, 2009 at 00:13 from IEEE Xplore. Restrictions apply.

	Channels				Routers		Latency					
Topology	N_C	b_C	N_{BC}	$N_{BC} \cdot b_C$	N_R	radix	H	T_R	T_C	T_{TC}	T_S	T_0
Crossbar	*64	*128	*64	8,192	1	64x64	1	10	n/a	0	4	14
Mesh	224	256	16	4,096	64	5x5	2-15	2	1	0	2	7-46
CMesh	48	512	8	4,096	16	8x8	1-7	2	2	0	1	3-25
Clos	128	128	64	8,192	24	8x8	3	2	2-10	0-1	4	14-32

Table 1: Example MTBw Network Configurations – Networks sized to support 128 b/cycle per tile under uniform random traffic. N_c = number of channels, b_C = bits/channel, N_{BC} = number of bisection channels, N_R = number of routers, H = number of routers along data paths, T_R = router latency, T_C = channel latency, T_{TC} = latency from tile to first router, T_S = serialization latency, T_0 = zero load latency. *Crossbar "channels" are the shared crossbar buses.

between neighboring tiles. Thus global electrical crossbars are unlikely choices for future manycore on-chip networks, despite the fact that they might be the easiest to program.

Two-dimensional mesh networks (Figure 1b) are popular in systems with more tiles due to their simplicity in terms of design, wire routing, and decentralized flow-control [3, 14]. Unfortunately, high hop counts result in long latencies and significant energy consumption in both routers and channels. Because network latency and throughput are critically dependent on application mapping, low-dimensional mesh networks also impact programmer productivity by requiring careful optimization of task and data placement.

Moving from low-dimensional to high-dimensional mesh networks (e.g., 4-ary 3-cubes) reduces the network diameter, but requires long channels when mapped to a planar substrate. Also, higher-radix routers are required, resulting in more area and higher router energy. Instead of adding network dimensions, researchers have proposed using concentration to help reduce hop count [1]. Figure 1c illustrates a two-dimensional mesh with a concentration factor of four (cmesh). One of the disadvantages of cmesh topologies is that, for the same theoretical throughput, channels are wider than an equivalent mesh topology as shown in Table 1. One option to improve channel utilization for shorter messages is to divide resources among multiple parallel cmesh networks with narrower channels. The cmesh topology should achieve similar throughput as a standard mesh with half the latency at the cost of longer channels and higher-radix routers. CMesh topologies still require careful application mappings for good performance.

Clos networks offer an interesting intermediate point between the high-radix, low-diameter crossbar topology and the low-radix, high-diameter mesh topology [4]. Figure 1d illustrates an 8-ary 3-stage Clos topology which reduces the hop count but requires longer point-to-point channels. Figure 2 shows one possible layout of this topology. Clos networks use many small routers and extensive path diversity. Although the specific Clos network shown here is reconfigurably non-blocking instead of strictly non-blocking, we can still minimize congestion with an appropriate routing algorithm (assuming the outputs are not oversubscribed). Unfortunately, Clos networks still require global point-to-point channels and, as with a crossbar, these global channels can be difficult to layout and have significant energy cost.

4. Photonic On-Chip Networks

Silicon photonics is a promising new technology which offers lower power, higher bandwidth density, and shorter latencies than electrical interconnects. Photonics is particularly effective for global interconnects and thus has the potential to enable scalable low-diameter on-chip networks, which should ease manycore parallel programming. In this section, we first introduce the underlying photonic technology before discussing the cost of implementing some of the global photonic crossbars found in the literature. We then introduce our own approach to implementing a photonic Clos network, and compare its cost to photonic crossbars.

4.1. Photonic Technology

Figure 3 illustrates the various components in a typical wavelength-division multiplexed (WDM) photonic link used for on-chip communication. Light from an off-chip two-wavelength (λ_1, λ_2) laser source is carried by an optical fiber and then coupled into an on-chip waveguide. The waveguide carries the light past a series of transmitters, each using a resonant ring modulator to imprint the data on the corresponding wavelength. Modulated light continues through the waveguide to the other side of the chip where each of the two receivers use a tuned resonant ring filter to "drop" the corresponding wavelength from the waveguide into a local photodetector. The photodetector turns absorbed light into current, which is sensed by the electrical receiver. Both 3D and monolithic integration approaches have been proposed in the past few years to implement silicon-photonic on-chip networks.

With 3D integration, a separate specialized die or layer is used for photonic devices. Devices can be implemented in monocrystalline silicon-on-insulator (SoI) dies with

Authorized licensed use limited to: MIT Libraries. Downloaded on December 1, 2009 at 00:13 from IEEE Xplore. Restrictions apply.

Design	Modulator and Driver Circuits DDE	FE	TTE	Receiver Circuits DDE	FE	TTE	ELP
Aggressive	20 fJ/bt	5 fJ/bt	16 fJ/bt/heater	20 fJ/bt	5 fJ/bt	16 fJ/bt/heater	3.3 W
Conservative	80 fJ/bt	10 fJ/bt	32 fJ/bt/heater	40 fJ/bt	20 fJ/bt	32 fJ/bt/heater	33 W

Table 2: Aggressive and Conservative Energy and Power Projections for Photonic Devices – fJ/bt = average energy per bit-time, DDE = Data-traffic dependent energy, FE = Fixed energy (clock, leakage), TTE = Thermal tuning energy (20K temperature range), ELP = Electrical laser power budget (30% laser effieiency).

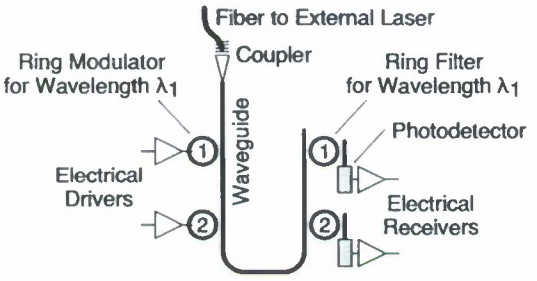


Figure 3: Photonic Components – Two point-to-point photonic links implemented with WDM.

Photonic device	Optical Loss (dB)
Optical Fiber (per cm)	0.5e-5
Coupler	1
Splitter	0.2
Non-linearity (at 30 mW)	1
Modulator Insertion	0 – 1
Waveguide (per cm)	0 – 5
Waveguide crossing	0.05
Filter through	1e-4 – 1e-2
Filter drop	1.5
Photodetector	0.1

Table 3: Optical Loss Ranges per Component

thick layer of buried oxide (BOX) [5], or in a separate layer of silicon nitride (SiN) deposited on top of the metal staek [2]. In this separate die or layer, customized processing steps can be used to optimize device performance. However, this customized processing approach increases the number of processing steps and hence manufacturing cost. In addition, the circuits required to interface the two chips can consume significant area and power.

With monolithic integration, photonic devices are designed using the existing process layers of a standard logic process. The photonic devices can be implemented in polysilicon on top of the shallow-trench isolation in a standard bulk CMOS process [9] or in monocrystalline silicon with advanced thin BOX SoI. Although monolithic integration may require some post-processing, its manufacturing cost can be lower than 3D integration. Monolithic integration decreases the area and energy required to interface electrical and photonic devices, but it requires active area for waveguides and other photonic devices.

Irrespective of the chosen integration methodology, WDM optical links have many similar optical loss components (see Table 3). Optical loss affects system design, as it sets the required optical laser power and correspondingly the electrical laser power (at a roughly 30% conversion efficiency). Along the optical critical path, some losses such as coupler loss, non-linearity, photodetector loss, and filter drop loss are relatively independent of the network layout, size, and topology. For the scope of this study, we will focus on the loss components which significantly impact the overall power budget as a function of the type, radix, and throughput of the network.

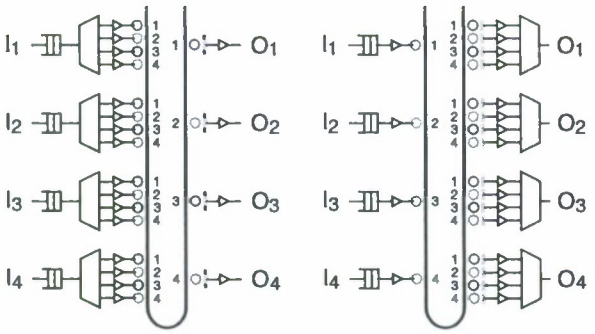
In addition to optical loss, ring filters and modulators have to be thermally tuned to maintain their resonance under on-die temperature variations. Monolithic integration gives the most optimistic ring heating efficiency of all approaches (due to in-plane heaters and air-undercut), estimated at 1 μW per ring per K.

Based on our analysis of various photonic technologies and integration approaches, we make the following assumptions. With double-ring filters and a 4 THz free-spectral range, up to 128 wavelengths modulated at 10 Gb/s can be placed on each waveguide (64 in each direction, interleaved to alleviate filter roll-off requirements and crosstalk). A non-linearity limit of 30 mW at 1 dB loss is assumed for the waveguides. The waveguides are single mode and a pitch of 4 μm minimizes the crosstalk between neighboring waveguides. The ring diameters are ≈10 μm. The latency of a global photonic link is assumed to be 3 cycles (1 cycle in flight and 1 cycle each for E/O and O/E conversion). For monolithic integration we assume a 5 μm separation between the photonic and electrical devices to maintain signal integrity, while for 3D integration the photonic devices are designed on a separate specialized layer. Table 2 shows our assumptions for the photonic link energy and electrical laser power.

4.2. Photonic Global Crossbar Networks

A global crossbar provides non-blocking all-to-all communication between its inputs and outputs in a single stage. Figure 4 shows two approaches for implementing a 4×4 photonic crossbar. Both schemes have multiple single-wavelength photonic channels carried on

Authorized licensed use limited to: MIT Libraries. Downloaded on December 1, 2009 at 00:13 from IEEE Xplore. Restrictions apply.



(a) DMXbar (b) CMXbar

Figure 4: Photonic 4x4 Crossbars – Both crossbars have four inputs (I_{1-4}), four outputs (O_{1-4}), and four channels which are wavelength division multiplexed onto the U-shaped waveguide. Number next to each ring indicates resonant wavelength. (a) distributed mux crossbar (DMXbar) with one channel per output, (b) centralized mux crossbar (CMXbar) with one channel per input.

a single waveguide using WDM. Crossbars with higher radix and/or greater channel bandwidths will require more wavelengths and more waveguides. Both examples require global arbitration to determine which input can send to which output. Various arbitration schemes are possible including electrical and photonic versions of centralized and distributed arbitration.

Figure 4a illustrates a distributed mux crossbar (DMXbar) where there is one channel per output and every input can modulate every output channel. As an example, if I_1 wants to send a message to O_3 it first arbitrates and then modulates wavelength λ_3. This light will experience four modulator insertion losses, 13 through losses, and one drop loss. Notice that although a DMXbar only needs one ring filter per output, it requires $O(nr^2)$ modulators where r is the crossbar radix and n is the number of wavelengths per port. For larger radix crossbars with wider channel bitwidths the number of modulators can significantly impact optical power, thermal tuning power, and area. For large distributed-mux crossbars this requires very aggressive photonic modulator device design. Vantrease et al. have proposed a global 64×64 photonic crossbar which is similar in spirit to the DMXbar scheme and requires about a million rings [15]. Their work uses a photonic token passing network to implement the required global arbitration.

Figure 4b illustrates an alternative approach called a centralized mux crossbar (CMXbar) where there is one channel per input and every output can listen to every input channel. As an example, if I_3 wants to send a message to O_1 it first arbitrates and then modulates wavelength λ_3. By default all ring filters at the receivers are slightly off-resonance so output O_1 receives the message by tuning in the ring filter for λ_3. This light will experience one modulator insertion loss, 13 through losses, three detuned receiver through losses, and one drop loss. If all ring filters were always tuned in, then wavelength λ_3 would have to be split among all the outputs even though only one output is ever going to actually receive the data. Although useful for broadcast, this would drastically increase the optical power. A CMXbar only needs one modulator per input (and so is less sensitive to modulator insertion loss), but it requires $O(nr^2)$ drop filters. As with the DMXbar, this can impact optical power, thermal tuning power, and area, and it necessitates aggressive reduction in the ring through loss. Additionally, tuning of the appropriate drop filter rings when receiving a message is done using charge injection, and this incurs a fixed overhead cost of 50 µW per tuned ring. Kırman et al. investigated a global bus-based architecture which is similar to the CMXbar scheme [7]. Nodes optically broadcast a request signal to all other nodes, and then a distributed arbitration scheme allows all nodes to agree on which receiver rings to tune in. Psota et al. have also proposed a CMXbar-like scheme which focuses on supporting global broadcast where all receivers are always tuned in [12].

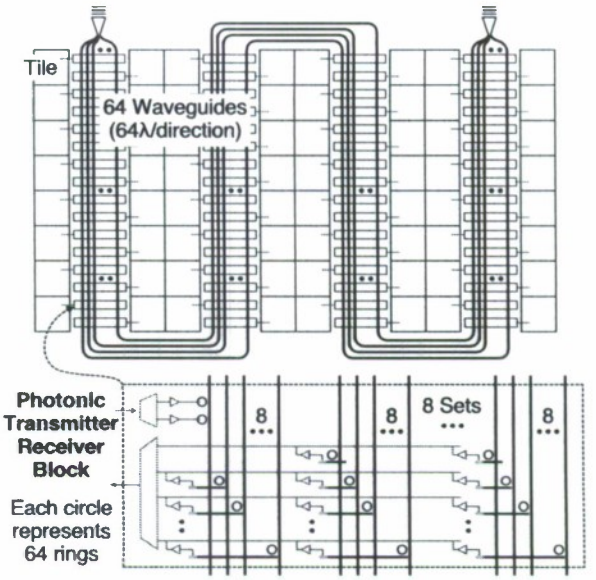


Figure 5: Serpentine Layout for 64x64 CMXbar – Electrical circuitry shown in red. 64 waveguides (8 sets of 8) are either routed between columns of tiles (monolithic integration) or over tiles (3D integration). One 128 b/cycle channel is mapped to each waveguide, with 64 λ going from left to right and 64 λ going from right to left. Each tile modulates a unique channel and every tile can receive from any channel.

Although Figure 4 shows two of the more common approaches proposed in the literature, there are other schemes which use a significantly different implementation. Zhou et al. describe an approach which replaces the U-shaped waveguide with a matrix of passive ring filters [17]. This approach still requires either multiple mod-

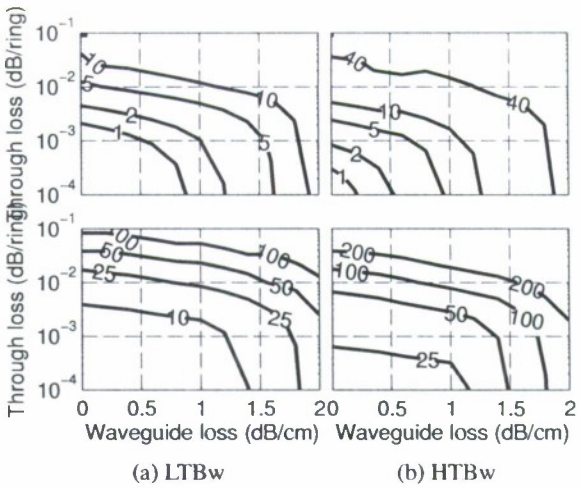
Authorized licensed use limited to: MIT Libraries. Downloaded on December 1, 2009 at 00:13 from IEEE Xplore. Restrictions apply.

ulators per input or multiple ring filters per output, but results in shorter waveguide lengths since all wavelengths do not need to pass by all tiles. Unfortunately, the matrix also increases the number of rings and waveguide crossings. Petracca et al. describe a crossbar implementation which leverages photonic switching elements that switch many wavelengths with a single ring resonator [10]. Their scheme requires an electrical control network to configure these photonic switching elements, and thus is best suited for transmitting very long messages which amortize configuration overhead. In this paper, we focus on the schemes illustrated in Figure 4 and leave a detailed comparison to more complicated crossbars for future work.

The DMXbar and CMXbar schemes can be extended to much larger systems in a variety of ways. A naive extension of the CMXbar scheme in Figure 4b is to layout a global loop around the chip with light always traveling in one direction. Unfortunately this layout has an optical critical path which would traverse the loop twice. Figure 5 shows a more efficient serpentine layout of the CMXbar scheme for our target system of 64 tiles. This crossbar has 128 b/cycle input ports which makes it suitable for a MTBw system (i.e., 128 b/cycle per tile under uniform random traffic). At a 5 GHz clock rate, each channel uses 64 λ (10 Gb/s/λ), and we need a total of 64 waveguides (1 waveguide/channel). An input can send light in either direction on the waveguides, which shortens the optical critical path but requires additional modulators per input.

The total power dissipated in the on-chip photonic network can be divided into two components. The first component consists of power dissipated in the photonic components, i.e., power at the laser source and the power dissipated in thermal tuning. The second part consists of electrical power dissipated in the modulator driver, receiver, and arbitration circuits. Here we quantify the first power component and then in Section 5 we provide a detailed analysis of the second power component.

The optical losses experienced in the various optical components and the desired network capacity determine the total optical power needed at the laser source. In the serpentine layout of a CMXbar, the waveguide and ring through loss are the dominant loss components, due to the long waveguides (9.5 cm) and large number of rings (128 modulator rings and $63 \times 64 = 4032$ filter rings) along each waveguide. Figure 6 shows two contour plots of the optical power required at the laser source for the LTBw and HTBw systems with a photonic CMXbar network. For a given value of waveguide loss and through loss per ring, the number of wavelengths per waveguide is the same for the two systems. However, the higher bandwidth system requires wider global buses which increases the optical power required at the laser source. As a result, the LTBw system can tolerate higher losses per component compared to the HTBw system for the same optical



(a) LTBw (b) HTBw

Figure 6: Laser Optical Power (W) (top row) and Percent Area (bottom row) for 64×64 CMXbar – Systems implemented with serpentine layout on 20×20 mm die.

	Global Crossbar		Clos	
System	Rings	Power	Rings	Power
LTBw	266 k	5.3 W	14 k	0.28 W
HTBw	1,000 k	21.3 W	57 k	1.14 W

Table 4: Thermal Power – Power required to thermally tune the rings in the network over a temperature range of 20K.

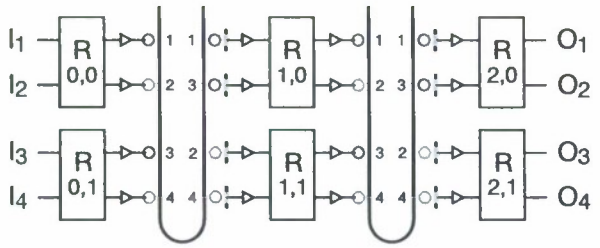
power budget.

Figure 6 shows contour plots of the percent area required for the optical devices for the LTBw and HTBw systems. The non-linearity limit affects the number of wavelengths that can be routed on each waveguide and hence the number of required waveguides, making photonic device area dependent on optical loss. As expected, the HTBw system requires increased photonic area for each loss combination. There is a lower limit on the area overhead which occurs when all of the wavelengths per waveguide are utilized. The minimum area for the LTBw and HTBw systems is 6% and 23%, respectively.

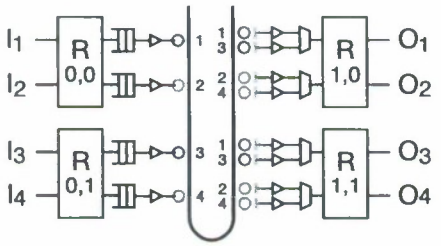
To calculate the required power for thermal tuning, we assume that under typical conditions the rings in the system would experience a temperature range of 20 K. Table 4 shows the power required for thermal tuning in the crossbar. Although each modulator and ring filter uses two cascaded rings, we assume that these two rings can share the same heater. The large number of rings in the crossbar significantly increases both thermal tuning and area overheads.

We can use a similar serpentine layout as the one shown in Figure 5 to implement a DMXbar. There would be one output tile per waveguide and there would be no need to tune or detune the drop filters. We would, however, require a large number of modulators per waveguide

Authorized licensed use limited to: MIT Libraries. Downloaded on December 1, 2009 at 00:13 from IEEE Xplore. Restrictions apply.



(a) Clos with Photonic Point-to-Point Channels



(b) Clos with Photonic Middle Routers

Figure 7: Photonic 2-ary 3-stage Clos Networks – Both networks have four inputs (I_{1-4}), four outputs (O_{1-4}), and six 2×2 routers ($R_{0-2,0-1}$). (a) four point-to-point photonic channels use WDM on each U-shaped waveguide. (b) the two middle routers ($R_{1,0-1}$) are implemented with photonic 2 × 2 CMXbars on a single U-shaped waveguide. Number next to each ring indicates resonant wavelength.

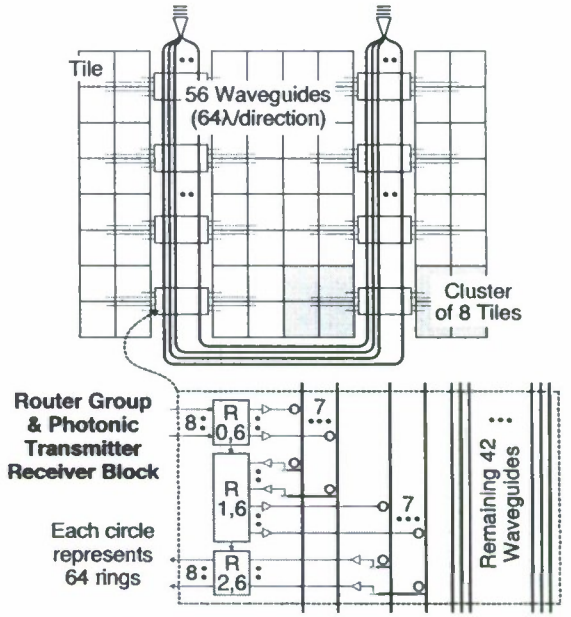


Figure 8: U-Shaped Layout for 8-ary 3-stage Clos – Electrical circuitry shown in red. 56 waveguides (8 sets of 7) are either routed between columns of tiles (monolithic integration) or over tiles (3D integration). Each of the 8 clusters (8 tiles per cluster) has electrical channels to its router group which contains one router per Clos stage. In the inset, the first set of 7 waveguides are used for channels (each 64 λ = 128 b/cycle) connecting to and from every other cluster. The second set of 7 waveguides are used for the second half of the Clos network. The remaining 42 waveguides are used for point-to-point channels between other clusters.

(63 × 64 = 4032) and modulator insertion loss would most likely dominate the optical power loss. For this topology to be feasible, novel modulators with close to 0 dB insertion loss need to be designed. The area for photonic devices and power dissipated in thermally tuning the rings would be similar to that in the CMXbar implementation.

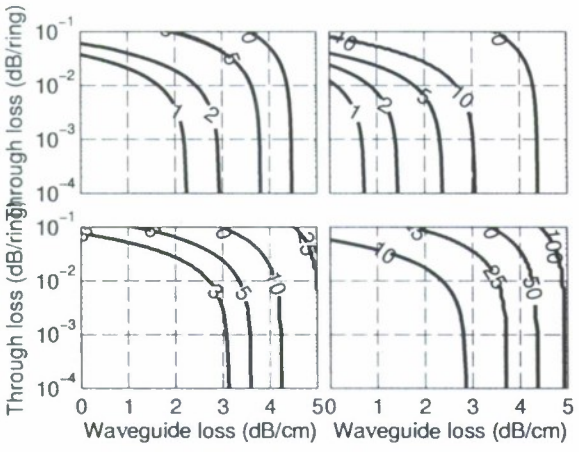
The large number of rings required for photonic crossbar implementations make monolithic integration impractical from an area perspective, and 3D integration is expensive due to the power cost of thermal tuning (even in the case when all the circuits of the inactive transmitters/receivers can be fully powered down). The actual cost of these crossbar networks will be even higher than indicated in this section since we have not accounted for arbitration overhead. These observations motivate our interest in photonic Clos networks which preserve much of the simplicity of the crossbar programming model, while significantly reducing area and power.

4.3. Photonic Clos Networks

As described in Section 3.2, a Clos network uses multiple stages of small routers to create a larger non-blocking all-to-all network. Figure 7 shows two approaches for implementing a 2-ary 3-stage Clos network. In Figure 7a, all of the Clos routers are implemented electrically and the inter-router channels are implemented with photonics. As an example, if input I_2 wants to communicate with output O_4 then it can use either middle router. If the routing algorithm chooses $R_{1,1}$, then the network will use wavelength λ_2 on the first waveguide to send the message to $R_{1,1}$ and wavelength λ_4 on the second waveguide to send the message to O_4. Figure 7b is logically the same topology, but each middle router is implemented with photonic CMXbar. The channels for both crossbars are multiplexed onto the same waveguide using WDM. Note that we still use electrical buffering and arbitration for these photonic middle routers. Using photonic instead of electrical middle routers removes one stage of EOE conversion and can potentially lower the dynamic power of the middle router crossbars, but at the cost of higher optical and thermal tuning power. Depending on photonic device losses, this tradeoff may be beneficial since for our target system the radix of the Clos routers (8×8) is relatively low. In this paper, we focus on the Clos with photonic point-to-point channels since it should have the lowest optical power, thermal tuning power, and area overhead.

As in the crossbar case, there are multiple ways to extend this smaller Clos network to larger systems. For a fair comparison, we keep the same packaging constraints (i.e., location of vertical couplers) and also try to use

Authorized licensed use limited to: MIT Libraries. Downloaded on December 1, 2009 at 00:13 from IEEE Xplore. Restrictions apply.



(a) LTBw (b) HTBw

Figure 9: Laser Optical Power (W) (top row) and Percent Area (bottom row) for 8-ary 3-stage Clos – Systems implemented with U-shaped layout on 20×20 mm die.

the light from the laser most efficiently. Figure 8 shows the U-shaped layout of the photonic Clos network in a MTBw system, which corresponds to 64 λ per channel. Each point-to-point photonic channel uses either forward or backward propagating wavelengths depending on the physical location of the source and destination clusters.

In a Clos network, the waveguide and ring through losses contribute significantly to the total optical loss but to a lesser extent than in a crossbar network, due to shorter waveguides and less rings along each waveguide. All the waveguides in the Clos network are roughly 2× shorter and with 20× less rings along each waveguide compared to a crossbar network. Figure 9 shows the optical power contours for the Clos network.

Although the number of optical channels in the Clos network is higher than in the crossbar network, the total number of rings (for same bandwidth) is significantly smaller since optical channels are point-to-point, resulting in significantly smaller tuning (Table 4) and area costs. The area overhead shown in Figure 9 is much smaller than for a crossbar due to shorter waveguides and smaller number of rings and is well suited for monolithic integration with a wider range of device losses. The lower limit on the area overhead is 2% and 8% for LTBw and HTBw, respectively.

Based on this design-space exploration we propose using the photonic Clos network for on-chip communication. Clos networks have lower area and thermal tuning costs and higher tolerance of photonic device losses as compared to global photonic crossbars. In the next section we compare this photonic Clos network with electrical implementations of mesh, cmesh, and Clos networks in terms of throughput, latency, and power.

5. Simulation Results

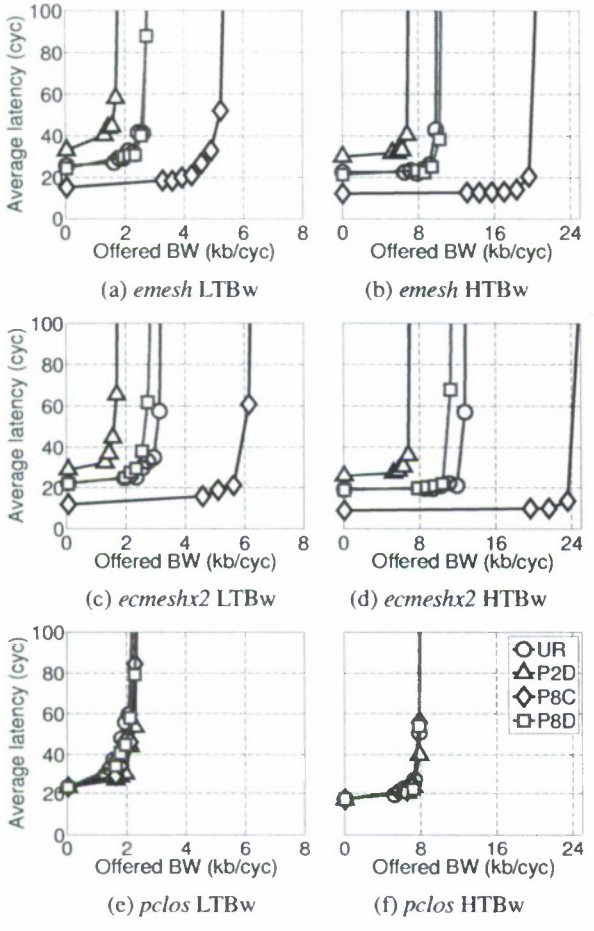
In this section, we use a detailed cycle-accurate microarchitectural simulator to study the performance and power of various electrical and photonic networks for a 64-tile system with 512 b messages. Our model includes pipeline latencies, router contention, flow control, and serialization overheads. Warm-up, measure, and drain phases of several thousand cycles and infinite source queues were used to accurately determine the latency at a given injection rate. Various events (e.g., channel utilization, queue accesses, arbitration) were counted during simulation and then multiplied by energy values derived from first-order gate-level models.

Our baseline includes three electrical networks: a 2D mesh (*emesh*), a mesh with a concentration factor of four (*ecmeshx2*), and an 8-ary 3-stage Clos (*eclos*). Because a single concentrated mesh would have channel bitwidths larger than our message size for some configurations, we implement two parallel emeshes with narrow channels and randomly interleave messages between them. We also study a photonic implementation of the Clos network (*pclos*) with aggressive (*pclos-a*) and conservative (*pclos-c*) photonic devices (see Table 2). We show results for LTBw and HTBw systems which correspond to ideal throughputs of 64 b/cycle and 256 b/cycle per tile for uniform random traffic. Our mesh networks use dimension-ordered routing, while our Clos networks use a randomized oblivious routing algorithm (i.e., randomly choosing the middle router). All networks use wormhole flow control.

We use synthetic traffic patterns based on a partitioned application model. Each traffic pattern has some number of logical partitions, and tiles randomly communicate only with other tiles that are in the same partition. These logical partitions are then mapped to physical tiles in either a co-located fashion (tiles within a partition are physically grouped together) or in a distributed fashion (tiles in a partition are distributed across the chip). We believe these partitioned traffic patterns capture the varying locality present in manycore programs. Although we studied various partition sizes and mappings, we focus on the following four representative patterns in this paper. A single global partition is identical to the standard uniform random traffic pattern (UR). The P8C pattern has eight partitions each with eight tiles optimally co-located together. The P8D pattern stripes these partitions across the chip. The P2D pattern has 32 partitions each with two tiles, and these two tiles are mapped to diagonally opposite quadrants of the chip.

Figure 10 shows the latency versus offered bandwidth for the LTBw and HTBw systems with different traffic patterns. In both *emesh* and *ecmeshx2*, the P8C traffic pattern requires only local communication and thus has higher performance. The P2D traffic pattern requires global communication which results in lower per-

Authorized licensed use limited to: MIT Libraries. Downloaded on December 1, 2009 at 00:13 from IEEE Xplore. Restrictions apply.



(a) *emesh* LTBw (b) *emesh* HTBw

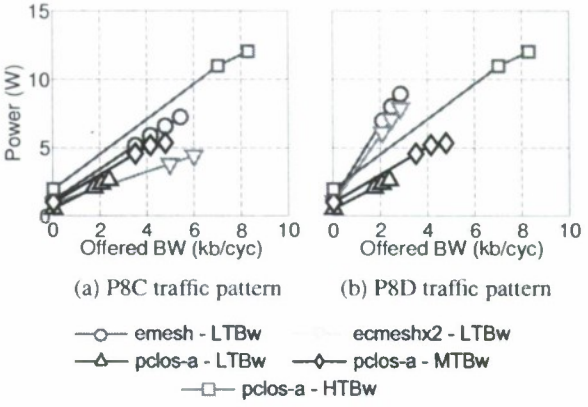
(c) *ecmeshx2* LTBw (d) *ecmeshx2* HTBw

(e) *pclos* LTBw (f) *pclos* HTBw

Figure 10: Latency vs. Offered Bandwidth – LTBw systems have a theoretical throughput of 64 b/cycle per tile for UR; corresponding for HTBw is 256 b/cycle.

formanee. On average, *ecmeshx2* saturates at higher bandwidths than *emesh* due to the path diversity provided by the two emesh networks, and has lower lateney due to lower average hop eount. Although not shown in Figure 10, the *eclos* network has similar saturation throughput to *pclos* but with higher averagc latency. Beeause *pclos* always distributes traffic randomly across its middle routers, it has uniform lateney and throughput across all traffie patterns. Note, however, that *pclos* performs better than *emesh* and *emeshx2* on global traffie patterns (e.g., P2D) and worse on local traffic patterns (e.g., P8C). If the *pclos* power eonsumption is low enough for the LTBw system then we should be able to inerease the size to a MTBw or HTBw system. A larger *pclos* network will hopefully have similar performanee and energy-effieieney for local traffie patterns as eompared to *emesh* and *ecmeshx2* and mueh better performanee and energy-cfficieney for global traffie patterns.

Figure 11 shows the power dissipation versus offered bandwidth for various network topologies with the P8C



(a) P8C traffic pattern (b) P8D traffic pattern

Figure 11: Power Dissipation vs. Offered Bandwidth – 3.3 W laser power not included for the *pclos-a* topology.

and P8D traffie patterns. In order to matеh the performanee of *ecmeshx2* LTBw system we need to use the *pclos-a* MTBw system whieh has slightly higher power for the P8C traffie pattern (local eommunieation) and mueh lower power for the P8D traffie pattern (global eommunieation) assuming we are at medium to high load. Laser power is not ineluded in Figure 11 whieh may be appropriate for systems primarily limited by the power density of the processor chip, but may not be appropriate for energy-eonstrained systems or for systems limited by the total power eonsumption of the motherboard.

Figure 12 shows the power breakdowns for various topologies and traffie patterns, for both LTBw and HTBw design points that ean support the desired offered bandwidth with lowest power. Compared to *emesh* and *ecmeshx2*, the *pclos-a* network provides eomparable performanee and low power dissipation for global traffie patterns, and eomparable performanee and power dissipation for loeal traffie patterns. The *pclos-a* network energy-effieieney inereases when sized for higher throughputs (higher utilization) due to statie laser power component. More importantly, the *pclos-a* network offers a global low-dimensional network with uniform performanee whieh should simplify manyeore parallel programming. The energy effieieney of *pclos* network might be further improved by investigating alternative implementations whieh use photonie middle switeh router as shown in Figure 7b.

It is important to note that with eonservative optieal teehnology projections, even in relatively simple optieal network like *pclos*, the required eleetrieal laser power is mueh larger than other eomponents, and the photonie network will usually eonsume higher power than the eleetrieal networks. This strong eoupling between overall network performanee, topology and underlying photonie eomponents underlines the need for a fully integrated vertieal design approaeh illustrated in this paper.

Authorized licensed use limited to: MIT Libraries. Downloaded on December 1, 2009 at 00:13 from IEEE Xplore. Restrictions apply.

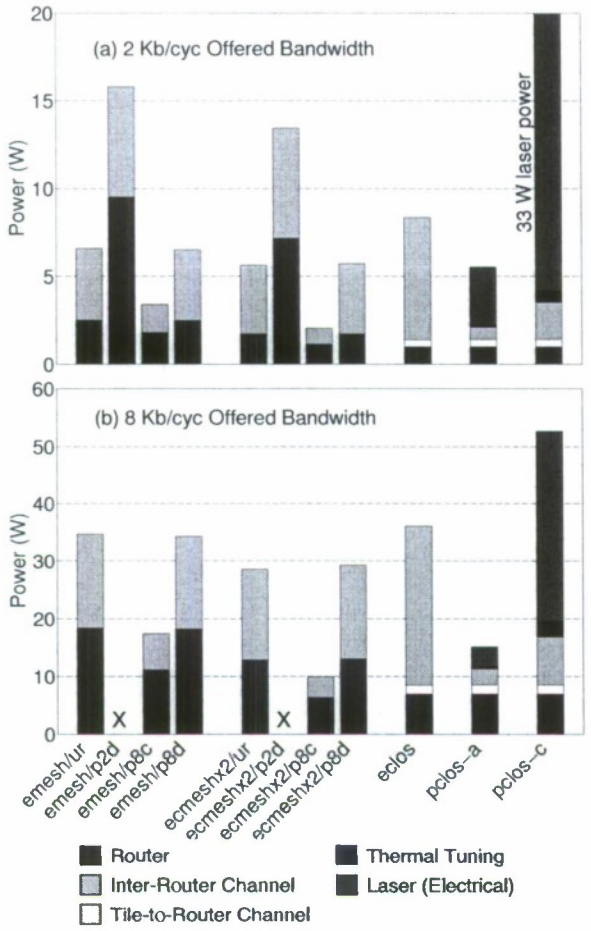


Figure 12: Dynamic Power Breakdown – Power of *eclos* and *pclos* did not vary significantly across traffic patterns. (a) LTBw systems at 2 kb/cycle offered bandwidth (except for *emesh/p2d* and *ecmeshx2/p2d* which saturated before 2 kb/cycle, HTBw system shown instead), (b) HTBw systems at 8 kb/cycle offered bandwidth (except for *emesh/p2d* and *ecmeshx2/p2d* which saturated before 8 kb/cycle).

6. Conclusion

We have proposed and evaluated a silicon-photonic Clos network for global on-chip communication. Since the Clos network uses point-to-point channels instead of the global shared channels found in crossbar networks, our photonic Clos implementations consume significantly less optical power, thermal tuning power, and area overhead, while imposing less aggressive loss requirements on photonic devices. Our simulations show that the resulting photonic Clos networks should provide higher energy-efficiency than electrical implementations of mesh and Clos networks with equivalent throughput. A unique feature of a photonic Clos network is that it provides uniformly low latency and uniformly high bandwidth regardless of traffic pattern, which helps reduce the programming challenge introduced by highly parallel systems.

Acknowledgments

This work was supported in part by Intel Corp. and DARPA awards W911NF-08-1-0134 and W911NF-08-1-0139.

References

[1] J. Balfour and W. J. Dally. Design tradeoffs for tiled CMP on-chip networks. *Int'l Conf. on Supercomputing*, 2006.

[2] T. Barwicz et al. Silicon photonics for compact, energy-efficient interconnects. *Journal of Optical Networking*, 6(1):63–73, 2007.

[3] S. Bell et al. TILE64 processor: A 64-core SoC with mesh interconnect. *Int'l Solid-State Circuits Conf.*, Feb. 2008.

[4] C. Clos. A study of non-blocking switching networks. *Bell System Technical Journal*, 32:406–424, 1953.

[5] C. Gunn. CMOS photonics for high-speed interconnects. *IEEE Micro*, 26(2):58–66, Mar./Apr. 2006.

[6] B. Kim and V. Stojanović. Characterization of equalized and repeated interconnects for NoC applications. *IEEE Design and Test of Computers*, 25(5):430–439, 2008.

[7] N. Kirman et al. Leveraging optical technology in future bus-based chip multiprocessors. *Int'l Symp. on Microarchitecture*, Dec. 2006.

[8] U. Nawathe et al. An 8-core 64-thread 64 b power-efficient SPARC SoC. *Int'l Solid-State Circuits Conf.*, Feb. 2007.

[9] J. Orcutt et al. Demonstration of an electronic photonic integrated circuit in a commercial scaled bulk CMOS process. *Conf. on Lasers and Electro-Optics*, May 2008.

[10] M. Petracca et al. Design exploration of optical interconnection networks for chip multiprocessors. *Symp. on High-Performance Interconnects*, Aug. 2008.

[11] D. Pham et al. The design and implementation of a first-generation CELL processor. *Int'l Solid-State Circuits Conf.*, Feb. 2005.

[12] J. Psota et al. ATAC: On-chip optical networks for multi-core processors. *Boston Area Architecture Workshop*, Jan. 2007.

[13] A. Shacham, K. Bergman, and L. Carloni. On the design of a photonic network-on-chip. *Int'l Symp. on Networks-on-Chip*, May 2007.

[14] S. Vangal et al. 80-tile 1.28 TFlops network-on-chip in 65 nm CMOS. *Int'l Solid-State Circuits Conf.*, Feb. 2007.

[15] D. Vantrease et al. Corona: System implications of emerging nanophotonic technology. *Int'l Symp. on Computer Architecture*, June 2008.

[16] W. Zhao and Y. Cao. New generation of predictive technology model for sub-45 nm early design exploration. *Trans. on Electron Devices*, 53(11):2816–2823, Nov. 2006.

[17] L. Zhou et al. Design and evaluation of an arbitration-free passive optical crossbar for on-chip interconnection networks. *Applied Physics A: Materials Science & Processing*, Feb. 2009.

Authorized licensed use limited to: MIT Libraries. Downloaded on December 1, 2009 at 00:13 from IEEE Xplore. Restrictions apply.

B. Design of Multi-Socket Core-to-DRAM Networks

Designing Multi-socket Systems Using Silicon Photonics

Scott Beamer, Krste Asanović
Department of Electrical Engineering and Computer Science
University of California, Berkeley, California
{sbeamer, krste}@eecs.berkeley.edu

Christopher Batten, Ajay Joshi, Vladimir Stojanović
Department of Electrical Engineering and Computer Science
Massachusetts Institute of Technology, Cambridge, Massachusetts
{cbatten, joshi, vlada}@mit.edu

ABSTRACT

Future single-board multi-socket systems may be unable to deliver the needed memory bandwidth electrically due to power limitations, which will hurt their ability to drive performance improvements. Energy efficient off-chip silicon photonics could be used to deliver the needed bandwidth, and it could be extended on-chip to create a relatively flat network topology. That flat network may make it possible to implement the same number of cores with a greater number of smaller dies for a cost advantage with negligible performance degradation.

Categories and Subject Descriptors: B.4.3 [Computer Systems Organization]: Processor Architectures[Parallel Architectures]

General Terms: Design, Economics, Performance

Keywords: Silicon Photonics, Multi-socket

1. INTRODUCTION

Given the difficulties of scaling uniprocessor performance further, most commercial microprocessor manufacturers have instead used increased transistor densities to integrate multiple processor cores on one die [1]. To deliver further performance improvements, multi-socket systems have been used to increase the computing power and memory capacity. These multi-socket systems will require increasing memory bandwidth to deliver realizable improvements in application performance. This bandwidth must come not only from connections to DRAM, but also from inter-socket links. Even if the bandwidth to these systems is not hampered by pin limitations, it will be restricted by power limitations from electrical off-chip signalling.

Silicon photonics could be used off-chip to solve this bandwidth problem, with its great potential for energy efficiency and bandwidth density. If photonics is used for the inter-socket links, it could also be extended on-chip closer to its destinations. In this work we present a scalable interconnect based on monolithically integrated silicon photonics that is able to harness the technology's potential to create an uniform network topology. With an approximately flat multi-socket interconnect, the penalty for communicating between sockets is reduced, which may enable potential cost benefits from implementing the same aggregate die area over a greater number of smaller dies.

Copyright is held by the author/owner(s).
ICS'09, June 8–12, 2009, Yorktown Heights, New York, USA.
ACM 978-1-60558-498-0/09/06.

2. SILICON PHOTONICS POTENTIAL

Silicon photonics has emerged in recent years as an appealing way to enable high bandwidths without excessive area or power requirements [3, 5, 6]. Due to the diversity in prospective photonic technologies, we selected a particular proposal for monolithically integrated silicon photonics [2, 4] to base our design and its evaluation on.

Since photonics uses light rather than electricity to transmit data, transmitted bits must undergo conversion at both ends (electro-optical and opto-electrical) which adds a constant latency and energy penalty. Because those penalties are constant, photonics excels over a distance due to greater amortization. The most compelling advantages for silicon photonics over forecasted electrical interconnects are its high bandwidth density and energy efficiency for off-chip communication. On-chip the selected technology performs well under the same metrics, but only if it travels a non-trivial distance. Using a coupler, it is possible to guide light from off-chip fibers onto on-chip waveguides without retransmission or modification. This enables seamless inter-chip links to be made, since if the constant conversion overhead is going to be paid for off-chip links, it makes sense to traverse the remaining on-chip portion optically as well since it is nearly free [2].

For the selected technology, laser light is generated in bulk off-chip and carried by fiber to splitters on-chip where it will be directed to the various links. Power is consumed by photonic links at the endpoints on-chip for signaling as well as the off-chip light source. Along the path from the transmitter to the receiver, there are various types of losses the signal incurs, and sufficient laser power must be applied to compensate. The *optical critical path* is the path with the most loss from the light source to the last receiver, and it dictates how much laser power the system will need.

3. ARCHITECTURE

For this research we target single-board multi-socket systems, and our design leverages the potential of photonics to produce a flat network. These boards could be connected together by another network to create an even larger system, but within a board a core sees uniform memory performance. Since electrical interconnects are advantageous over short distances, we electrically join groups of cores (4–16) into *clusters* by shared L2 caches. These clusters are connected by dedicated photonic links to every memory controller (Figure 1). Fully connected networks are often avoided because of their quadratic growth in resource consumption, but the

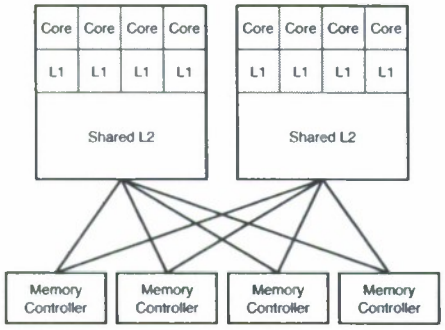


Figure 1: Topology for two clusters of four cores with four memory controllers

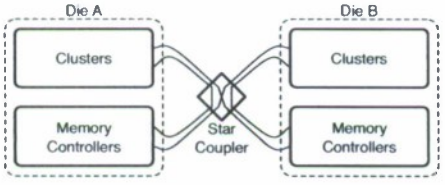


Figure 2: Logical view of a two die system

energy efficiency and the bandwidth density advantages of silicon photonics make it tolerable for a small system. A memory controller may actually communicate with multiple DRAM channels, but from the point of view of the network, it is simply a point of arbitration for access to that memory. The links between the DRAM modules and the memory controllers are electrical because of the challenges involved with changing the DRAM interface, but future work could benefit greatly if these connections were photonic.

The simple network topology was not only chosen to make a flat network, but to also enable a single die design to be used in varying quantities to make a scalable range of systems. In this glueless system, a cluster's memory bandwidth is uniformly spread across all the memory controllers, so in the maximum supported system size there is one direct channel between each cluster and each memory controller. For systems with less populated sockets, each cluster will get the same total bandwidth, but it will have multiple channels to each memory controller. To enable this flexibility, the actual connections between memory controllers and clusters are done off chip (Figure 2) so the changes necessary for systems of different sizes are localized to small off-chip components. To simplify the packaging and assembly of all the point-to-point connections, off-chip fibers are grouped into ribbons, which connect to a *star coupler*. The star coupler is a passive device that connects two groups of ribbons such that each ribbon has at least one fiber directly coupled with a fiber from every ribbon in the other group. Our design template is general enough that it is able to scale down to smaller dies while maintaining the same topology and nearly identical performance.

4. INCENTIVES FOR DISINTEGRATION

Using a greater number of smaller dies to implement the same silicon area could have cost advantages. Smaller dies should benefit from higher yield rates and increased tolerance to process variation, since they could be binned on finer granularities. A single reusable design will also have a higher sales volume, which will reduce non-recurring engineering (NRE) costs. This disintegration is made worthwhile by photonics, because otherwise it will increase the number of electrical pins and power spent on the interconnect. For our design, smaller dies will allow the system to be more spread out, which will reduce the power density and make it easier to electrically attach DRAM. Fixed costs per die (testing, packaging, and assembly) will cause penalties for using dies that are too small, but the optimum die size for cost may be smaller than current commercial designs.

5. RESULTS

Using our candidate technology, we evaluated the general design while varying the die size (16–256 cores/die) and the maximum supported system size (64–1024 cores). To scale to higher core counts will require a multi-hop network. The layout of each design was optimized to reduce the optical critical path loss because laser power can be the majority power consumer of a photonic interconnect. The area taken by the on-chip interconnect was always less than 10%, and the latency stayed roughly constant since the network topology stayed the same. Interestingly, for the range of designs explored, independent of the total numbers of cores, systems with a modest number of dies (4–8) had the lowest optical power.

6. CONCLUSION

Silicon photonics provides an appealing way to supply the bandwidth needed to drive multi-socket systems, and a range of scalable designs capable of supporting up to 1024 cores with uniform memory bandwidth was presented. In a relatively flat network like the one presented, silicon photonics sufficiently reduces the barrier to going off-chip such that future die sizes may be chosen by what is most cost efficient rather than what is most reasonable to manufacture.

7. REFERENCES

[1] K. Asanovic et al. The landscape of parallel computing research: A view from berkeley. Technical report, U.C. Berkeley, 2006.

[2] C. Batten et al. Building manycore processor-to-dram networks with monolithic silicon photonics. *Symposium on High-Performance Interconnects*, Jan 2008.

[3] N. Kirman et al. Leveraging optical technology in future bus-based chip multiprocessors. *IEEE Micro*, 27(6), Jan 2006.

[4] J. Orcutt et al. Demonstration of an electronic photonic integrated circuit in a commercial scaled bulk cmos process. *Conf. on Lasers and Electro-Optics*, 2008.

[5] A. Shacham et al. Photonic noc for dma communications in chip multiprocessors. *Symposium on High-Performance Interconnects*, 15, Jan 2007.

[6] D. Vantrease et al. Corona: System implications of emerging nanophotonic technology. *ISCA*, Jan 2008.

Designing Multisocket Systems with Silicon Photonics

by Scott Beamer

Research Project

Submitted to the Department of Electrical Engineering and Computer Sciences, University of California at Berkeley, in partial satisfaction of the requirements for the degree of **Master of Science, Plan II**.

Approval for the Report and Comprehensive Examination:

Committee:

Professor Krste Asanović
Research Advisor

(Date)

* * * * * * *

Professor David A. Patterson
Second Reader

(Date)

Abstract

To fuel an increasing need for parallel performance, system designers have resulted to using multiple sockets to provide more hardware parallelism. These multisocket systems have limited off-chip bandwidth due to their electrical interconnect which is both power and pin limited. Current systems often use of a *Non-Uniform Memory Architecture* (NUMA) to get the most system memory bandwidth from limited off-chip bandwidth. A NUMA system complicates the work of a performance programmer or operating system, because they must maintain data locality to maintain performance.

Silicon photonics is an emerging technology that promises great off-chip bandwidth density and energy efficiency when compared to electrical signaling. With this abundance of bandwidth, it will be possible to build a relatively flat, high bandwidth memory interconnect. Because this interconnect has uniform bandwidth, NUMA optimizations will be unnecessary, which increases performance programmer productivity.

If the penalties to making a multi-socket system are negated by the use of silicon photonics, there is less incentive to integrate, and economic incentives to disintegrate. In this thesis, we present this scalable and coherent multi-socket design along with discussing the tradeoffs facing an architect when incorporating silicon photonics technology.

Chapter 1

Introduction

Given the difficulties of scaling uniprocessor performance further, most commercial microprocessor manufacturers have instead used increased transistor densities to integrate multiple processor cores on one die [1]. These manycore systems will require increasing memory bandwidth at reasonable energy consumption if they are to deliver improvements in application performance. Otherwise these systems may be grossly underutilized [27].

When the desired number of cores cannot fit on a die that is economical to manufacture, they are spread across multiple sockets. To feed many cores spread across multiple sockets will require even more memory bandwidth. Each socket will have its own attached DRAM, but in a shared memory machine it must be made accessible to the other sockets within the system. This interconnect must have an on-chip portion that connects all of the cores within a socket in addition to an off-chip portion that connects all the sockets within the system.

Current multisocket systems often have their off-chip bandwidth constrained by power and pin limitations [14, 18, 23]. As more cores are integrated into a die within a socket, they will need even more bandwidth, and this bottleneck will become more troublesome as it is unlikely off-chip electrical bandwidth will be able to keep up. The energy required to send a bit between sockets is not scaling down very quickly because the sockets are not getting much closer physically, and the materials used for traces is not getting significantly less resistive or capacitive. Even if off-chip electrical signaling becomes sufficiently more energy efficient, pin bandwidth could become the next limiting factor. Off-chip signaling rates and die sizes are not growing fast enough to provide enough pin bandwidth to meet the growing demand.

A socket's limited off-chip bandwidth must be divided up between links to its own locally attached DRAM and inter-socket links to reach remote DRAM attached to other sockets (Figure 1.1). If all of the bandwidth is allocated to the locally attached DRAM, the system will have the maximum memory bandwidth possible, but it will be disjoint. In contrast, if all of the bandwidth is allocated to the inter-socket links, the system will have no memory bandwidth but great inter-core bandwidth. If the two are balanced uniformly such that each socket receives an equal amount of bandwidth from every part of memory (remote or local) the system will have a *Uniform Memory Architecture* (UMA), and if they are balanced non-uniformly, the system will have a *Non-Uniform Memory Architecture* (NUMA).

Systems trying to get the most system memory bandwidth while coping with off-chip bandwidth scarcity will be pushed towards a NUMA design. This is true independent of the off-chip network topology, because each inter-socket link occupies bandwidth at two sockets, while a link to DRAM

only occupies bandwidth at one socket. Any bandwidth taken away from the inter-socket links, can be turned into twice the bandwidth for the links to DRAM. This encourages system designers to skew the bandwidth allocations in favor of locally attached DRAM instead of reaching other sockets, to maximize system memory bandwidth.

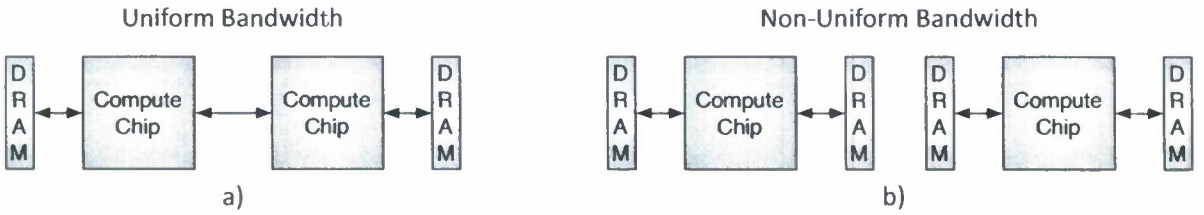


Figure 1.1: Motivation for NUMA

A NUMA design imposes additional complexity on the performance programmer, as it is crucial that data is co-located with the computation using it. This careful mapping is yet another optimization performance programmers must consider [27], but if the memory system was flat (uniform) it would be unnecessary, increasing their productivity. Some multiprogrammed workloads, such as virtual machines running within a datacenter, will also benefit from the scheduling flexibility that bandwidth uniformity provides. When scheduling jobs, a job could be run on the first available core independent of where the data it needs resides. Furthermore, some workloads exhibit poor spatial locality so it is difficult to spread the data across sockets effectively. If a new technology provided an abundance of bandwidth, it would be worthwhile to allocate it uniformly to increase programmer productivity and make the system more flexible.

In this work, we leverage silicon photonics to design high and uniform bandwidth multi-socket memory interconnects. We present a general network design that can be used to make systems of varying sizes, and to provide shared memory which makes the system more usable, we discuss how to reasonably implement coherency on top of the network. Because of the nature of the design, it has much less incentive to integrate, which opens the door to chip disintegration for cost savings. Overall, multi-socket interconnects are an interesting place to explore applications of current research in silicon photonics because of its emphasis on off-chip communication.

Chapter 2

Photonic Technology Introduction

Over the last few decades, the scale at which optical technology has been adopted for communication has been steadily decreasing. Optical communication was first used for long distance telecommunications, because its high endpoint costs were amortized over very long links. As processing technologies have improved, the cost (delay, space, energy, dollars ...) of the endpoints have decreased, which in turn has decreased the distance at which optical communication is advantageous. Continued technology advances along with increased integration have enabled silicon photonics, which decreases the feasible distance down to the inter-chip and even intra-chip level.

2.1 Technology Overview

In recent years, silicon photonics has been shown to be an increasingly desirable technology for system interconnects because of its potential for higher bandwidth density, greater energy efficiency, and lower latency. The technology is still immature with many competing implementation proposals, so projected performance on these important metrics varies significantly. To ground the results of our study, we select a particular monolithically integrated silicon photonics technology [4], but the overall approach should be applicable to the other current proposals because much of it is based on general technology insights.

Figure 2.1 shows a basic link is comprised of: a light source, a modulator, a waveguide, and a photodetector. The modulator encodes the signal by absorbing or not absorbing light as it passes by it through the silicon waveguide. At the other end of the waveguide, the photodetector senses the changes in light and decodes the signal. The electro-optical and opto-electrical conversions at the endpoints introduce a latency and energy cost that needs to be amortized beyond a minimum distance to be advantageous to electrical.

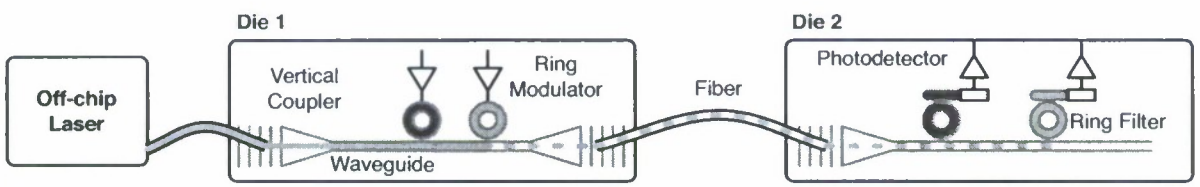


Figure 2.1: An inter-socket photonic link

The selected technology provides Dense Wave Division Multiplexing (DWDM) which contributes to its high bandwidth density (bits/second/μm). DWDM allows light from different wavelengths to share the same waveguide with minimal interference, which allows multiple logical links to share the same physical media without time multiplexing. This is enabled by putting rings which resonate with a narrow frequency of light onto the waveguide, such that when the light resonates with a ring, it is pulled off the waveguide into the ring. We can use these rings along with charge injection to make a ring modulator [11, 20, 21]. Applying a charge to a ring shifts the ring's resonant frequency so a particular wavelength can be absorbed or not absorbed to modulate the light.

A filter can also be made by using these resonant rings [21, 26], and the selected technology uses two cascaded rings to get additional frequency selectivity (Double Ring Filter). Since the photodetectors are sensitive to a wide range of light frequencies, a double ring filter is placed between the photodetector and the waveguide so only the correct wavelength gets through the filter and strikes the photodetector. These resonant rings are sensitive to a variety of environmental factors and manufacturing variations, but these can be combated by thermally tuning the rings with in-plane heaters.

The selected technology is monolithically integrated, and it utilizes a current CMOS manufacturing process which makes it much more realizable since it leverages a great deal of manufacturing hardware investment and knowledge. Other photonic proposals may be better suited for transmitting light, but they use materials or steps not currently part of a standard CMOS process making them more cost prohibitive to implement [3, 11, 15].

The light used by the system is generated by an off-chip laser because conventional CMOS processes are poorly suited for laser fabrication. This light is brought on chip through a fiber and then a coupler into the waveguide. On-chip light travels through poly-Si, which can be made into a usable waveguide (Figure 2.2) by placing it on top of shallow trench isolation (STI) and etching an air gap underneath it [10]. The air gap helps to improve the cladding on the bottom of the waveguide, because the STI is too thin on its own. The air gap does take up silicon area, so when possible multiple waveguides should share one to amortize the overhead. A great advantage of photonics is that once the signal has been encoded optically, that light can be guided through through couplers and a fiber to another chip's waveguide without retransmission (Figure 2.1), enabling links that operate seamlessly across long distances.

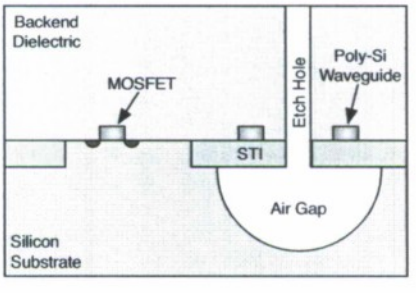


Figure 2.2: Cross section of an on-chip waveguide

2.2 Performance

Looking forward to when this silicon photonic proposal might be fully realizable, we compare it against a projected optimally repeated electric wire in a $22nm$ process and Tables 2.1, 2.2, and 2.3 give a summary of the comparison. Based on preliminary results and device projections, the silicon photonic proposal assumes a signaling rate of 10Gbps (faster could be possible) and squeezes in 64 wavelengths per direction [21], meaning a single link (fiber or waveguide) has 80GB/s of bidirectional bandwidth.

Table 2.1: Approximate energy costs per bit

Quantity	Electric (fJ)	Photonic (fJ)	Ratio
On-Chip Model	$50\frac{fJ}{mm}$	150[1]	
Off-Chip Model	5000	150	
Local On-Chip Wire (1 μm)	0.05	150	0.00033
Intermediate On-Chip Wire ($1mm$)	50	150	0.33
Global On-Chip Wire ($10mm$)	500	150	3.33
Off-Chip Trace ($40mm$)	5000	150	33.33
Chip-to-Chip Link ($40mm$ off-chip, $10mm$ on-chip)	5500	150	33.67

Table 2.2: Approximate latency costs per bit

Quantity	Electric (ps)	Photonic (ps)	Ratio
On-Chip Model	$100\frac{ps}{mm}$	$200 + 10\frac{ps}{mm}$	
Off-Chip Model	$50 + 5\frac{ps}{mm}$	$200 + 5\frac{ps}{mm}$	
Local On-Chip Wire (1 μn)	0.1	200.01	0.0005
Intermediate On-Chip Wire ($1mm$)	100	210	0.48
Global On-Chip Wire ($10mm$)	1000	300	3.33
Off-Chip Trace ($40mm$)	250	400	0.63
Chip-to-Chip Link ($40mm$ off-chip, $10mm$ on-chip)	1250	700	1.79

Table 2.3: Approximate bandwidth densities per bit. Photonic values sum the bandwidth of both directions

	Electric (Gb/s/ μm)	Photonic (Gb/s/ μn)	Ratio
On-Chip	5	320	64.0
Off-Chip	0.2	26	130.0

[1] $100\frac{fJ}{b}$ (modulator) + $50\frac{fJ}{b}$ (receiver) + $80uW$ (power to thermally tune rings) + optical power

2.2.1 Power

Energy efficiency ($\frac{bits/sec}{W} = \frac{bits}{J}$), especially off-chip, has been listed as one of the strongest advantages of the selected photonic technology. It is important to fully explore the three ways it expends power:

- *Encoding/Decoding* power is consumed at the endpoints and it includes electrical circuits to serialize/deserialize the signal from the native system clock to the transmission rate as well as the power consumed by charge injection to modulate the signal. This power is insensitive to distance, is mostly dynamic, and the values quoted in Table 2.1 are for 100% utilization.

- *Light Generation* power is burned by the laser to produce the light used for communication. This power is constant, independent of utilization. It is difficult to dynamically adjust laser power. To generate laser light more efficiently, the same laser is used for multiple links, so unless all of the links are inactive, it is hard to scale back. It is important to note that the light generation power is the amount of electrical power required to produce the *laser power* (light intensity) the system needs. Light generation power is often overlooked, and most of the prior work has not added it to the power total with the justification that it is off chip and thus does not contribute to power density hotspots on the processor [25]. Keeping with convention, for most of this work laser power will be presented separately, because laser light generation is an orthogonal area of research, so converting it to electrical power might be misleading. However, when calculating the total power for a system, a conservative estimate of future laser efficiency of 25% is used. This power is strongly dependent on how much loss the path has, and Section 2.3 will present more details about this.

- *Thermal Tuning* power is burned up by heaters to control the ring's resonant frequency for process variation. The observed sensitivity is $1\mu W$/ring/K and the needed control range is 20K, so each ring will burn $20\mu W$.

In summary, using a silicon photonic link purely on-chip will not be significantly advantageous with regards to energy, unless it travels a substantial distance ($> 3mm$), however off-chip it could be more than an order of magnitude more efficient.

2.2.2 Latency

Most of the latency for a silicon photonic link is at the endpoints, since light propagates rapidly. The endpoint latency is a consequence of serializing and deserializing the data from the native clock rate to the transmission rate of 10 Gbps. Table 2.2 shows that photonics only has lower latency than electrical beyond $2.2mm$ on-chip. As mentioned earlier, the photonic links can go inter-chip without retransmission, so in those cases the latency gap between electric and photonic is further reduced.

2.2.3 Area

On-chip waveguides are larger than wires and they have a wider pitch. The air gaps makes the waveguides effectively wider because no circuits can be placed over them. Even though waveguides take up more area than wires, there is so much more bandwidth per waveguide from DWDM and bidirectional communication that it still obtains a large bandwidth density advantage (Table 2.3).

Off-chip this advantage becomes more significant because they have comparable pitches, with the same data rates, but a single fiber contains 64 links in each direction while an electrical pin only implements a single link in one direction.

2.3 Laser Power

Every optical component introduces some amount of loss to the signal, increasing the laser power needed to ensure sufficient light reaches every photodetector. As mentioned, in 2.2.1, light generation power is significant, and it is directly proportional to laser power. We define the *optical critical path* as the path with the greatest loss between the light source and the last photodetector. Along the optical critical path, the laser power required to overcome losses tends to grow exponentially rather than linearly, so a reasonable design can quickly become unreasonable when scaled up. The network layout and size can contribute greatly to loss, so careful physical layout design is essential to save power.

Using Figure 2.3, we can trace out an example optical critical path and show where the losses come from. Table 2.4 is included to give sense of the relative magnitudes, since the absolute values could change as the technology matures. The optical critical path starts at the laser, and ends at the last photodetector (the one attached to the filter for the green wavelength). Traveling any distance, the light experiences some loss, which is negligible for off-chip fibers and significant for on-chip waveguides. To go from from off-chip to on-chip or vice versa, the light travels through a coupler, which incurs loss substantial enough that links which span more than two chips may be untenable. Once the light has been brought on-chip, it typically is fanned out through splitters to make all of the needed links. When the waveguide crosses another waveguide, it also incurs loss because all waveguides are routed in the same plane with this technology. Crossing losses can be significant, because often multiple waveguides are routed parallel to each other, so a crossing actually results in many crossings.

There are a variety of losses caused by the resonant rings. When light passes by a filter tuned for another wavelength, it experiences *through loss* (Filter to through node). When it passes through the intended filter and reaches the photodetector, it experiences *drop loss* (Filter to drop node). *Modulator insertion loss* is incurred when a wavelength of light passes by a modulator tuned for that frequency that is currently inactive.

Another important consideration is the non-linearity limit imposed by the Poly-Si waveguide. As the combined power of the light inside a waveguide grows, there is a non-linear increase in the amount of light that escapes. To combat that loss, more laser power is used which results in even more loss, so its best to keep the total power for a waveguide within reasonable limits. Normally how many wavelengths can be put into a waveguide is set by the frequency selectivity of the photonic components used, but the number of wavelengths used per waveguide may also be set by the path loss which determines the power required per wavelength and thus the number of wavelengths that can fit under the non-linearity limit. The designs presented later in this study were made to have low loss, and they should be able to carry 64 wavelengths per direction without issue.

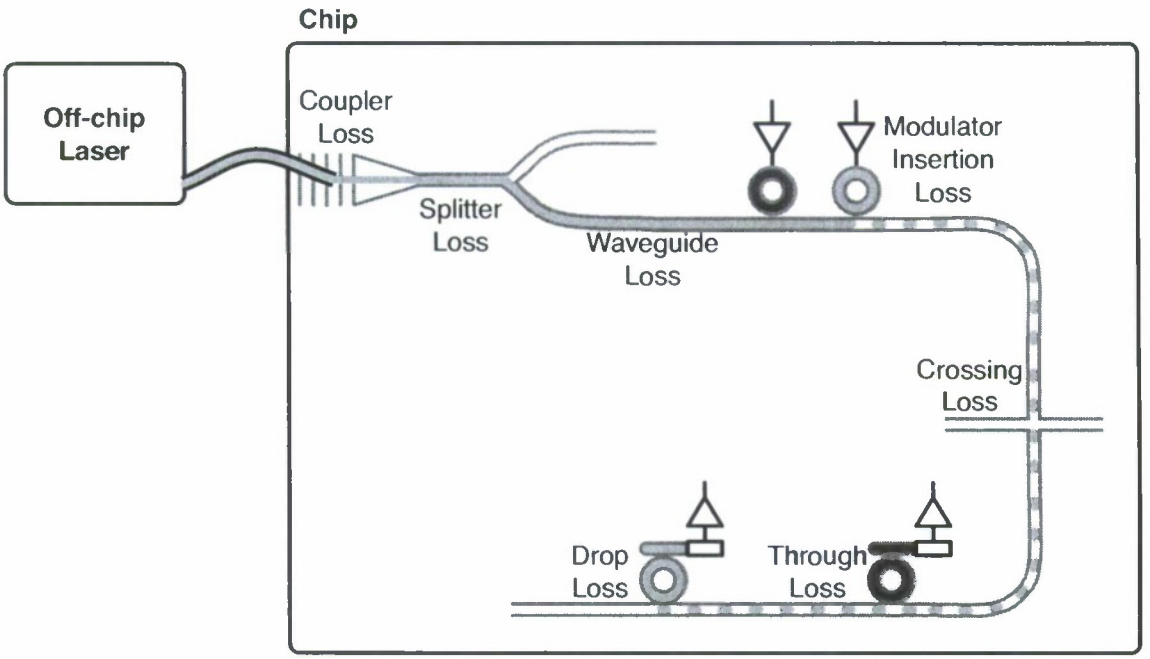


Figure 2.3: Photonic link with with losses labelled for the green wavelength

Table 2.4: Optical Power Costs [4]

Component	Loss (dB)
Coupler	1.0
Splitter	0.2
Non-Linearity	1.0
Filter (to through node)	0.01
Modulator Insertion	0.5
Waveguide Crossing	0.05
Waveguide (per cm)	1.0
Optical Fiber (per cm)	0.000005
Filter (to drop node)	1.5
Photodetector	0.1

2.4 Design Implications

As shown by Tables 2.1 & 2.3, the selected photonics technology can provide a tremendous amount of off-chip bandwidth, because of its energy efficiency and bandwidth density advantages. Replacing the electrical inter-socket links with photonic ones will enable much more bandwidth to each socket. Used in conjunction with an electrical on-chip network, it could still result in dramatically higher system bandwidth.

Even though entirely on-chip photonic links do not hold much of an advantage over electrical on-chip links, if photonics is used for the off-chip network, it makes sense to continue seamlessly on-chip because the conversions costs will have already been paid. By using these seamless links, off-chip networks and on-chip networks are flattened into one domain. To get the most from this flat network will require co-design of the on-chip and off-chip networks.

In this thesis, the connection between a memory controller and a DRAM module is assumed to be electrical. Future work could investigate a photonic link between a memory controller and DRAM, and doing so should not change the results of this study.

Chapter 3

Design of a Photonic Multisocket System

Section 2 shows that silicon photonics has great potential, and in this section we present a network designed to take full advantage of it. When designing a system known to be multi-socket, it is important to consider the off-chip network in addition to the on-chip network, and co-designing the on-chip and off-chip networks makes best use of seamless photonic links.

3.1 System Assumptions

To provide structure for the rest of this study, we make some assumptions about the target system. There are a variety of architectures that could take advantage of the transistor gains from Moore's law, but to achieve high computational throughput on a workload without high arithmetic intensity, they will all require high memory bandwidth. For this work, we envision a system comprised of many simple in-order cores, but some of the higher level results should still be applicable to other architectures.

To ground our design with real numbers (Table 3.1), we assume in a 22 nm process with 400 mm^2 of silicon, it will be possible to fit 256 cores running at 2.5GHz [4]. Each of these cores will include 4-way SIMD with Fused Multiply Accumulate (FMAC), giving the the system a total of 5 TFLOPS of peak performance. The amount of memory bandwidth needed to adequately supply this system will depend on the arithmetic intensity of the target workload, but the frequently desired ratio of one byte of memory bandwidth per one flop will support many desired workloads, which will equate to 5 TBps of memory bandwidth for the system [27]. This bandwidth will be supplied by 16 memory controllers, and each of these memory controllers may be attached to multiple physical DRAM channels, but from the point of view of the rest of the system, each memory controller is a single endpoint of arbitration and contention. We also assume that this system will be implemented over four sockets, so each socket will have one quarter of the cores and memory controllers. We assume a shared-memory system, where photonics is used to connect processor to memory controllers, not cores to cores.

	Baseline Socket	Max Configuration
Sockets	1	4
Cores	64	256
Clock Rate	2.5 GHz	2.5 GHz
Total Silicon Area	100 mm^2	400 mm^2
Memory Bandwidth	1.25 TBps	5 TBps
Memory Controllers	4	16

Table 3.1: Target system assumptions

3.2 Topology Insights

A network designer must balance the needs of the target workload with what the technology allows. The assumed workload for this system will need high bandwidth to feed many functional units, but this bandwidth must be provided uniformly (equally by all memory controllers) to simplify programming and to increase portability. Memory latency must be keep moderately low since the cores are mostly scalar, so they are incapable of cheaply tolerating too much memory latency. By Little's Law, the amount of data in flight is proportional to the product of latency and bandwidth. If the memory latency is increased, additional area will need to be dedicated to holding and tracking the increased amount of data in flight, which will make the simple cores more expensive.

A low-diameter, high-radix network will achieve these goals, and it will map well to the selected silicon photonics technology proposal. Low-diameter networks are known for low latency due to their low hop count, as well as having more uniform latency because there is less variance in path length [8]. This low hop count also results in more uniform bandwidth because there are less hops for links to get congested by other traffic on the network. To reach the same number of endpoints, a lower-diameter network must compensate with a higher radix. With a constant bandwidth per endpoint, increases in radix result in decreased bandwidth per link, which can be problematic as it will increase the serialization latency.

A common challenge with implementing low-diameter, high-radix networks in electrical technologies is that the links tend to become longer, and as a consequence, consume a significant amount of power. The selected photonic technology, however, is mostly distance insensitive with respect to latency and power. Another challenge with implementing these global links is that when mapped to a physical substrate, the bisection bandwidth required is high. This can be troublesome to route off-chip, but fortunately the selected photonic technology provides great off-chip bandwidth density. In contrast, if this network was implemented electrically, the bisection bandwidth would be constrained by the electrical pins, limiting the total network bandwidth. This would encourage the network designer to use a higher-diameter, lower-radix network to reduce the demand for bisection bandwidth which will also reduce the demand for off-chip bandwidth, at the price of longer and less uniform latencies and less uniform bandwidths.

Our design takes the low-diameter, high-radix network to the extreme, by using a simple fully-connected network (Figure 3.1a) as a starting point. Each network endpoint (core or memory controller), will have a high-radix switch with a photonic link for every possible endpoint. A single photonic hop minimizes latency while maximizing bandwidth uniformity. A one-hop topology will become a limiting factor as the design is scaled up to higher numbers of endpoints, since it will also increase the radix. Increasing the radix will hurt performance because the serialization latency will

grow as the links get narrower, and the power and area for the electrical switch will grow as its radix does. For the intended design scale of a single-board, compelling systems might be possible utilizing the selected photonic technology without taking up an unreasonable amount of area or power.

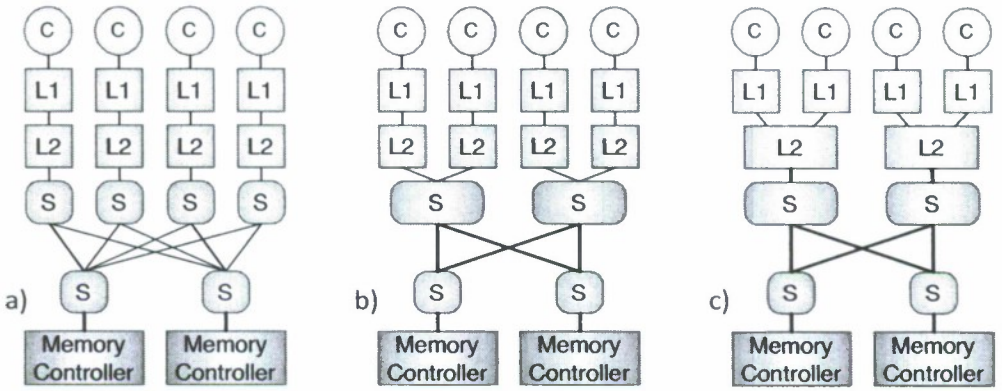


Figure 3.1: Topological benefits of concentration. a: Fully-connected network b: Fully-connected network with core concentration c: Fully-connected network with core concentration done by a shared cache

3.2.1 Concentration

Taking the simple initial design of a fully connected topology between individual cores and memory controllers and scaling it to meet the target system parameters will result in poor performance. The effective radix is high because there are so many memory controllers and cores, making each core–memory controller link so narrow (for the target system: $\frac{1}{16}^{th}$ of a core's bandwidth) that the serialization latency is significant. It is also statistically harder for a simple core to have enough memory request parallelism to keep all of those links busy simultaneously, leaving many of them underutilized. Low utilization is worrisome because static power constitutes a large fraction of a photonic link's power, but this can be avoided by using *concentration* to share links to increase utilization [8].

By grouping cores into *clusters* (Figure 3.1b), concentration widens the links to the memory controllers, which drastically cuts down on serialization latency. Since each cluster contains multiple cores, within in a cluster it is statistically easier to generate enough memory request parallelism to obtain higher utilization. Concentration combines the switches and links at the core side of the network to reduce the effective radix of the network. This has the desired effect of improving serialization latency, but it could also be used to build larger networks with the same serialization latency.

Since the cores within a cluster will be physically near each other as they share the same photonic cluster–memory links, they could also share their last level cache (Figure 3.1c). There are architectural benefits of sharing a cache, and current caches have been built with 8-way sharing [17]. These short links between cores and caches, and caches and the local switch should electrical, since it is too short of a distance for photonics to be advantageous. For the rest of our designs we

assume 8-core clusters, which obtains the benefits of concentration without overly burdening the cluster interconnect, but clusters of 2–16 cores should also be feasible.

3.2.2 Off-Chip Connections

With multi-socket systems it is desirable if the same chip can be used by simply varying the number connected together (even if only powers of two), because it will increase the volume of that part, lowering its cost. This scalable reusability is difficult to obtain while providing the goal of uniform memory bandwidth. As shown in Figure 3.2a, if the connections between clusters and memory controllers are made on-chip, that bandwidth is fixed because we want to reuse the same chip in all systems. Using that chip to build systems with a variable number of sockets populated will require some bandwidth (on-chip or off-chip) be turned off to keep the bandwidth allocation between the memory controllers on-chip and the memory controllers in other sockets uniform. If every connection is made off-chip (Figure 3.2b), the bandwidth allocations can be changed off-chip without modifying the chip.

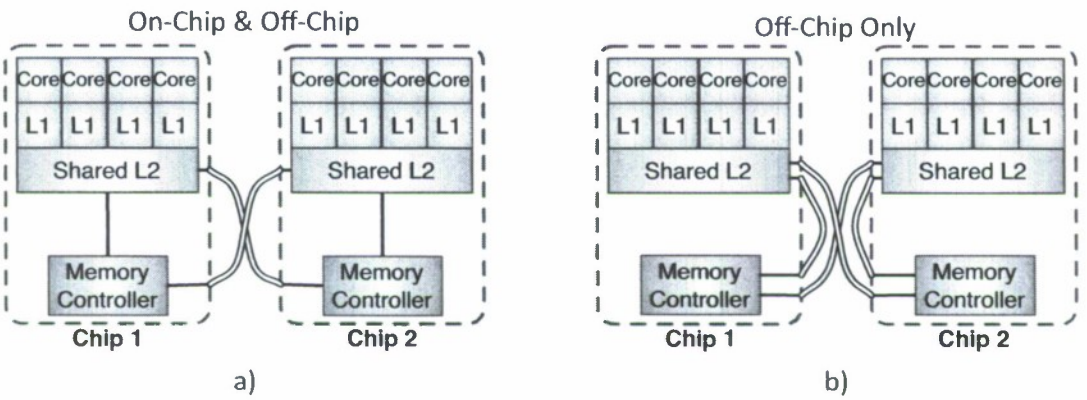


Figure 3.2: Topological benefits of all connections off-chip

To implement this, each cluster will have enough links to support the maximum number of memory controllers in the largest possible system, and the memory controllers will have enough links to support the maximum number of clusters in the largest possible system. In a fully populated system, all of these links will be connected one to one. If the system has only half of its sockets populated, there will be two links between each cluster and each memory controller. These links could be ganged together to make a single logical link of twice the bandwidth, or they could be kept separate to allow for greater memory request parallelism. In the case of a single socket system, the off-chip fibers are looped back.

It might seem that routing all traffic off-chip is wasteful when some of it could be done purely on-chip, but with photonics this penalty is greatly reduced. Most of the latency and on-chip energy cost of a photonic link is at the endpoints, so whether the link is purely on-chip or not only affects optical power. Depending on what the optical critical path loss is, this change in optical power may be truly negligible. This is in contrast to electrical off-chip links which consume sufficiently more energy and area such that an efficient design will never send data off-chip unless forced. Taking advantage of the off-chip bandwidth density, energy efficiency, and distance insensitivity of photonic

links, for the flexibility it provides and for the uniformity it maintains, the benefits of making all connections off-chip outweigh the small light generation power increase.

3.3 Packaging

To package the topology into a physical design will require more innovation. Because all of the cluster–memory controller connections are off-chip, each chip will have two types of fibers: those originating at clusters and those originating at memory controllers. Somehow off-chip, all of these fibers must be appropriately attached. To keep the fibers more organized, they can be grouped into *ribbons*, which simplifies assembly. As the number of sockets in the system grows, the number of ribbons that must be attached could be become unreasonable, because the topology is fully connected, so each socket must a have a ribbon to every socket (including itself). Figure 3.3a shows this for the four socket case.

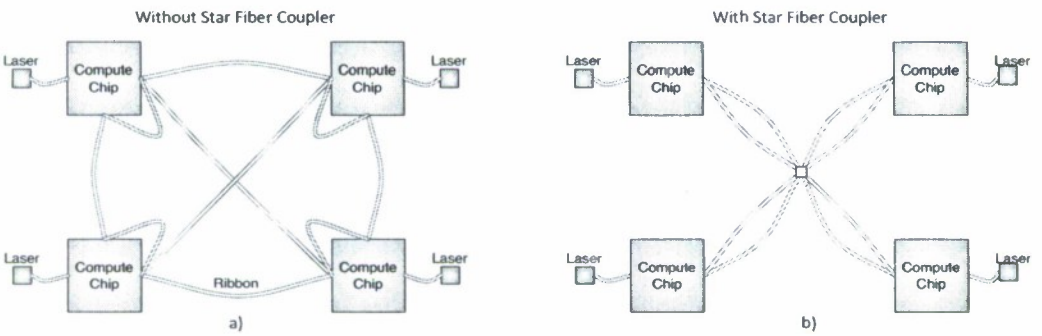


Figure 3.3: Comparison of with/without star fiber coupler

A *star fiber coupler* provides the needed all-to-all connectivity while greatly simplifying the fiber routing (Figure 3.3b). The star fiber coupler acts as a hub chip, so independent of system size, each socket only needs to attach two ribbons (one from its clusters and one from its memory controllers) to the coupler, and it will create the all-to-all connections. As shown in Figure 3.4, all of the cluster ribbons attach to one side of the coupler, and all of the memory controller ribbons attach to the other side. The ribbons from both sides come in orthogonal to each other so each ribbon crosses every other ribbon. In the example shown, four ribbons of four fibers come in each side, so effectively it is as if there is a fiber between every socket including itself (one fiber gets looped back).

The star fiber coupler can be generalized to support cases when there are more fibers than sockets or when multiple fibers are destined for each socket. It is a completely passive device, whose only purpose is to precisely hold ribbons such that their fibers can be efficiently coupled. The star fiber coupler should be comparably inexpensive, and along with some of the ribbons, are the only things to change between different system sizes.

To lay the system out on a board, the compute dies that contain the clusters and memory controllers are placed around the star fiber coupler as shown in Figure 3.5a. Each of the compute dies is surrounded by its own locally attached DRAM to reduce the distance for the electrical links between them. The memory controllers are evenly spaced around the edge of the die to provide

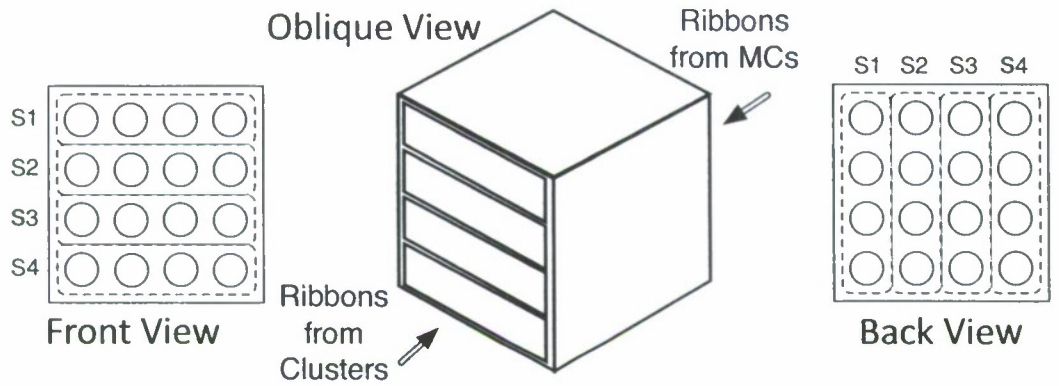


Figure 3.4: Schematic of star fiber coupler

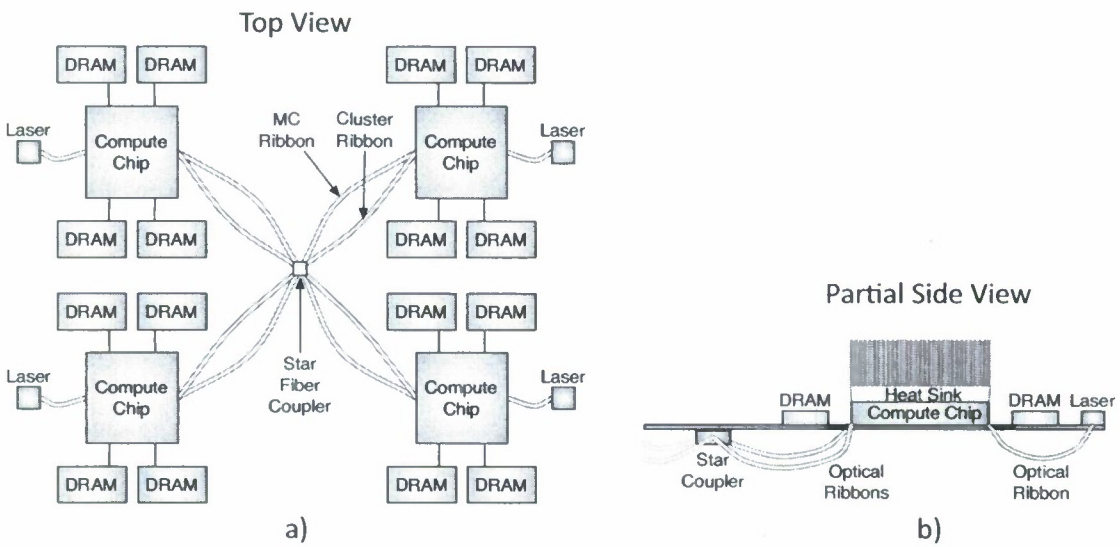


Figure 3.5: System layout

the easiest exposure for wiring to DRAM. By using photonics for inter-socket links, only a small amount of area needs to be dedicated to the fibers, leaving the rest of the pin area for connecting to DRAM or attaching to power and ground. The ribbons are attached only at the endpoints by vertical couplers and the ribbons will float freely beneath the board (Figure 3.5b), so they can avoid the heat sinks of the compute dies. A more dense board layout might reduce ribbon lengths, but it could significantly complicate the much more costly electrical signaling to DRAM or increase the power density. Extra distance in the ribbon is tolerable since the additional optical power loss and the increase in delay are both negligible.

3.4 Die Layout

The layout of the photonic components on-chip is crucial because it can greatly affect the optical power. Without careful design, the loss along the optical critical path quickly becomes so great that the laser power becomes unreasonable. Essentially the designer's job is to take all of the logical links, map those to wavelengths, and then map those to appropriate waveguides. The following sections highlight the optimizations used to make an efficient layout, such as the 64 core die layout in Figure 3.6.

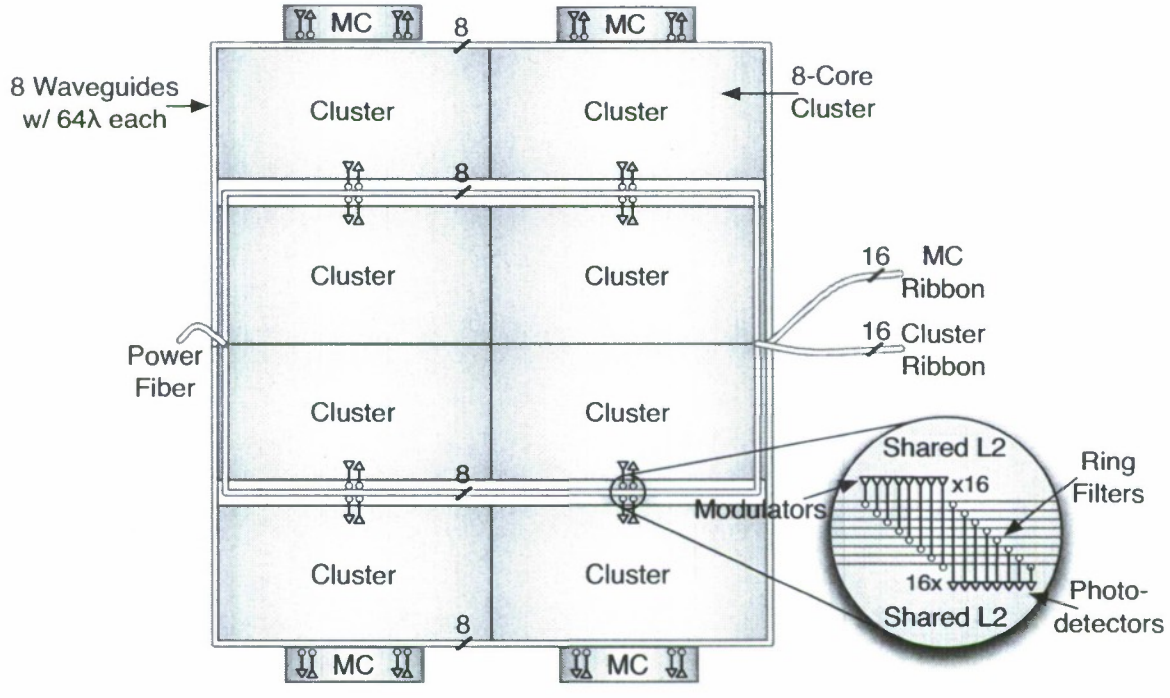


Figure 3.6: Die layout for 64 core die designed to support a 256 core system

3.4.1 Nested-U Waveguide Layout

Laying out the waveguides in a *nested-U* configuration as shown in Figure 3.6 can help combat two sometimes avoidable sources of optical loss: waveguide length and crossing loss. By bringing the power fiber in on one side and the inter-socket ribbons on the other, the waveguide distance is minimized while still reaching all the needed endpoints. The nested-U layout guarantees the the waveguide distance is less than or equal to the length of the chip plus the width of the chip. A single crossing doesn't contribute too much loss, but quite often waveguides are routed in parallel so a crossing will intersect multiple waveguides and then the losses multiply. Nesting the waveguides removes all crossings, since they always go around each other.

3.4.2 Cluster Striping Across Waveguides

With the nested-U waveguide layout, a waveguide from the power fiber to the inter-socket ribbon actually passes by more than one cluster. To load a waveguide with wavelengths from only one cluster exclusively is wasteful, because later on those wavelengths will need to be mixed for the inter-socket fibers. Striping a cluster's wavelengths across all the waveguides that pass by reduces the need to mix wavelengths later on.

In the example in Figure 3.6, eight waveguides pass four clusters. If each cluster put all of its wavelengths on two waveguides, somehow the wavelengths will need to be shuffled around such that they map appropriately to the four fibers that go between each socket. A device like the one presented in Section 4.2.1 could accomplish the needed mixing, but with striping it is often unnecessary.

3.5 Evaluation

As mentioned in Section 3.1, the network presented is designed to support 4 sockets of 64 cores, for a total of 256 cores. As an early evaluation of feasibility, we analyze its interconnect performance using conservative overestimates (Table 3.2). The system at theoretical peak can provide each core with the desired 1 byte : FLOP, for a total of 5 TBps of memory bandwidth.

Table 3.2: Overestimates for Quad Socket Interconnect for 256 Cores Total

Quantity	Value
Total Power	9.77W
Latency	1ns
Area (per socket)	4.2 mm^2

Figure 3.7 shows a breakdown of where the power is consumed in the interconnect. For our analysis we use the impractical 100% utilization to show what the peak power could be. With 0% utilization, the encoding/decoding power will scale down to about half of what it is at peak utilization, but the rest of the interconnect power is static and will not change based on activity. The encoding/decoding power is directly related to the number of photonic endpoints, and with a constant number of cores it will scale directly with the amount of offered bandwidth per core. Light generation power is burned in off chip lasers so it adds to the system wall power but not

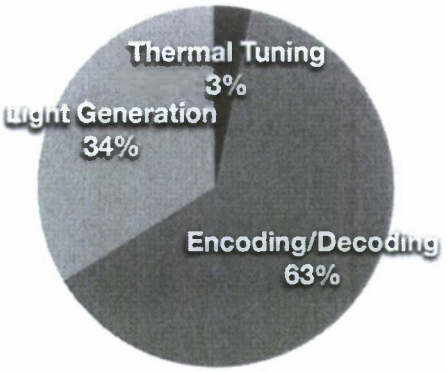


Figure 3.7: Breakdown of network power consumption for 256 core system with 64 cores per die

to the compute die's power density. For comparison, we converted laser power to light generation power assuming a conservative laser efficiency of 25%. Thermal tuning power is set by the number of rings which is also directly proportional to the number of photonic endpoints. This power is continuously burned, but it is not a large overall contributor (Figure 3.7).

The latency will depend on how far apart the sockets are placed, but if the off-chip fiber is under 11cm, the latency will be under 1ns (2-3 cycles for out target clock of 2.5GHz). This latency is actually quite good when it is put in context with other steps in memory operations such as L2 cache access latencies or DRAM access latencies. For our target system, the serialization latency will be 16 cycles for a 64B cache line, so in 18-19 cycles a cache line could move from a memory controller to a cluster's cache.

The area was grossly overestimated to give generous gaps between photonic components and the transistors around them. The area in Table 3.2 is per die, and in our target system each is $100\,\mathrm{mm}^2$, so that is only 4.2% overhead.

Integrating a new technology will have its costs, and they will have to be justified by dramatic performance improvements. Fortunately, the photonic network presented here will make some other parts of the system cheaper or easier to design. For example, since all inter-socket communication will be carried over fibers, this will dramatically reduce the number of traces that need to be routed on the printed circuit boards (PCB). This will make the PCB easier to design, cheaper to manufacture, and it will leave more space for other signals. Routing all inter-socket data through fibers will also mean that there will be less pins coming out of the socket, allowing for a smaller and cheaper package to be used. The increase in delay or energy for an increase in fiber length is marginal, which will give the system designer more flexibility in where they place sockets. In summary, using photonics simplifies much of the rest of the system, which will hopefully lessen the cost of adopting a new technology.

Chapter 4

Die Size Exploration

The design presented in Section 3 can be generalized to handle greater numbers of cores or even different die sizes. Since all connectivity is off-chip and we leverage the distance insensitivity of photonics, there is less motivation to integrate and an economic incentive to disintegrate.

4.1 Incentives for Disintegration

Disintegration may be able to reduce the cost of the system (relative to another made with the same template). Smaller dies could reduce costs by:

- *Increased yield.* Figure 4.1 shows the relative costs of manufacturing 400 mm^2 of silicon as one whole die or many smaller dies. Although the combined cost of the smaller dies is always cheaper due to increased yield, most of the gain can be had by splitting the die four ways to get a $\approx 3\times$ cost advantage. Figure 4.1 is from a simple model [12] that only takes into account parameters for area and defect densities. In the real world there will also be fixed costs (packaging, assembly, and test) per die that will make the systems with the smallest dies less desirable, but there still will be significant advantage to using multiple moderately smaller dies rather than a single large die.

- *Better binning.* Since the dies are smaller they can be binned on a finer granularity to reduce the impact of process variation. Within a small die, the probability of there being high process variance is reduced, allowing a greater number of high performance dies to be sold.

- *Greater design reuse.* As mentioned previously, being able to use the same die in systems of different sizes allows for greater amortization of non-recurring engineering (NRE) costs over the increased manufacturing volume. Smaller dies are easier to reuse because they support a greater variety of system sizes.

Smaller dies could also make system design easier. With smaller dies, possibly spread farther apart, the board-level power density of the system is reduced making cooling easier. It will be even easier to interface to adjacent electrically connected DRAM with smaller dies, since there will be less memory surrounding each die.

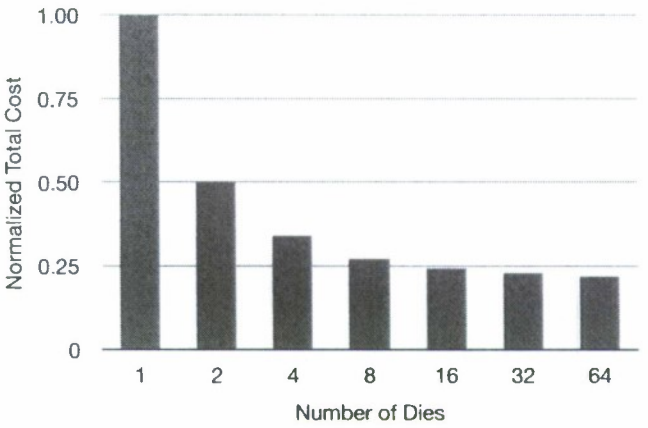


Figure 4.1: Relative total costs for 400 mm^2 of total silicon area

4.2 Scaling the Design

The design presented in Section 3 can scale to some other sizes, but in this section we describe two further photonic structures that increase the feasible range of designs.

4.2.1 Mixer

DWDM allows multiple logical links to share the same waveguide, but when a link needs to cross to another waveguide a *mixer* can be used. For each waveguide on one side, all of its wavelengths are evenly and disjointly distributed across the waveguides on the other side. It is a bidirectional component, and Figure 4.2 shows a simplified case, where two wavelengths from one waveguide are separated onto two waveguides. It is possible to extend this design to handle multiple waveguides per input group, so a NxN mixer (k wide) mixes N groups of k waveguides each. With this abstract notation, a wide range of components can be classified as mixers, and many of these special cases have already appeared in various other photonic designs [6, 24, 28].

Figure 4.2: Simplified 2x2 mixer (1 wide). Only one waveguide's wavelengths shown for simplicity.

To take the system from Section 3.4 from 256 cores total to 1024 cores total (still 64 cores per socket) will require two 2x2 mixers (8-wide) placed where the inter-socket ribbons attach to the on-chip waveguides (Figure 4.3). To reach 1024 cores with 64 core dies (100 mm^2), there are 16

sockets so each inter-socket link has only 1 fiber. Scaling in this manner keeps the bandwidth per core constant, but it does come from a greater number of memory controllers. The input groups to the mixers correspond to the groups of waveguides on the die. Without striping, the mixers would have to have more input groups.

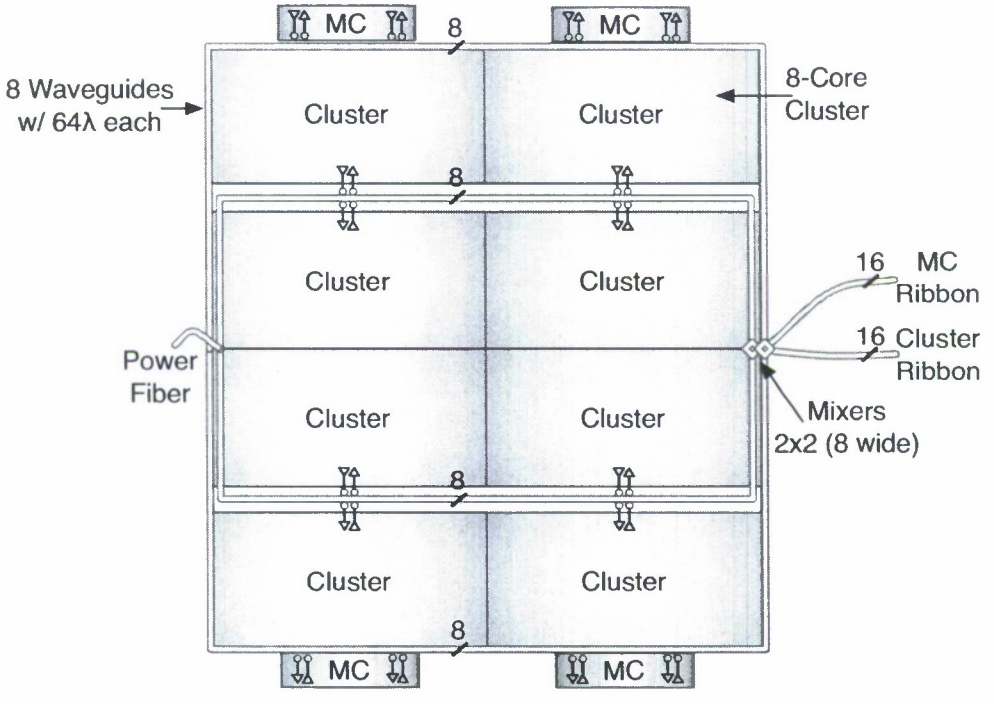


Figure 4.3: Die layout for 64 core die designed to support a 1024 core system

4.2.2 Add-Drop Multiplexer

When there are more dies in the system than waveguides on a die (this often happens with small dies), an *add-drop multiplexer* (ADM) can be used to fan out the wavelengths of one waveguide onto multiple underloaded waveguides. This component is bidirectional, so from one direction it looks like a splitter but from the other it looks like an aggregator. As shown in Figure 4.4 this can be done without crossings. Alternatively the die layout could simply under-fill the waveguides on-chip, but this wastes area and the optical loss through the ADM is low.

4.3 Evaluation

Using the generalized design template, we explore a range of possible systems with maximum capacities of 64 – 1024 cores built from 4 – 64 dies. We keep the cluster size the same (8 cores), the ratio of memory controllers to cores the same (1:16), and the same core density (0.64 cores/ mm^2). Table 4.1 shows what additional components (mixers or ADMs) are required to build systems of various sizes. There are tradeoffs when designing the base building block (die) for the system,

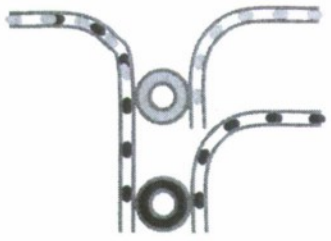


Figure 4.4: Simplified Add-Drop Multiplexer with two wavelengths per waveguide

both in terms of how big it is and how many other blocks it expects. If the maximum system size is designed too small, it will not be able to scale to larger systems without penalties, but if it is designed too large, the functionality needed for larger systems will waste area and raise cost when used in smaller systems. Some places where this tradeoff becomes apparent are: off-chip bandwidth, off-chip link organization, and coherency. For our particular family of designs, how populated the system is does not noticeably affect performance once the die size and the maximum system size have been set.

Table 4.1: Additional component requirements (mixers and ADMs) per die to scale the system size. The fanout degree for the ADM is on the top line, while the mixer degree is the bottom line.

	64 cores/system	128 cores/system	256 cores/system	512 cores/system	1024 cores/system
16 cores/die		2x	4x	8x	16x
32 cores/die			2x2 (4 wide)	2x 2x2 (4 wide)	4x 2x2 (4 wide)
64 cores/die					2x2 (8 wide)
128 cores/die					
256 cores/die					

4.3.1 Power

Since we keep the bandwidth per core constant, the encoding/decoding power remains constant at 24mW per core, whether we scale the number of cores or the number of dies to implement them (Figure 4.5). Since some of the higher core count designs use additional rings for filters in the interconnect (in ADM's and mixers), they will have slightly higher thermal tuning power but it is still negligible. These additional components will have a much larger impact with increased loss on the optical critical path which will increase the light generation power significantly.

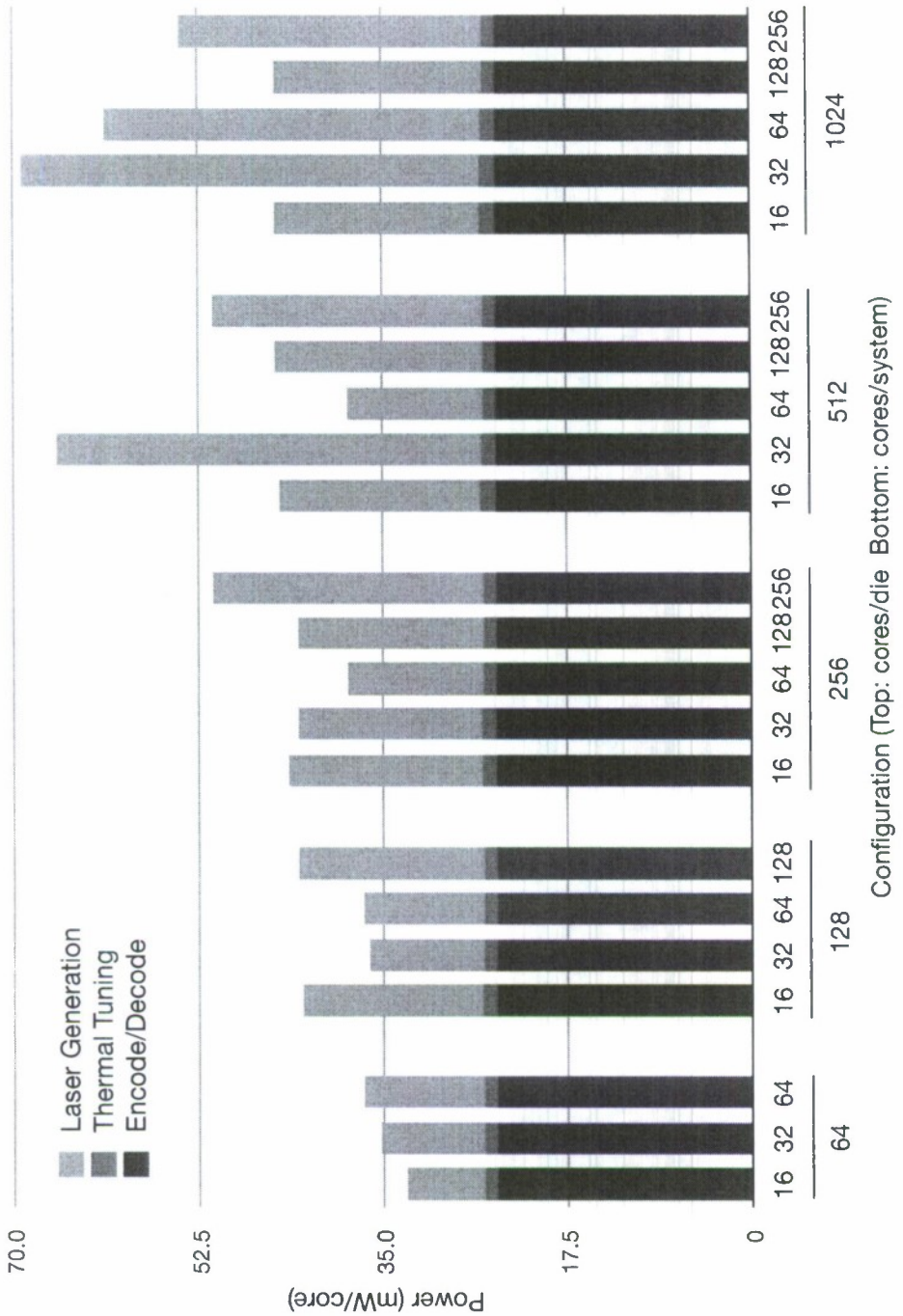


Figure 4.5: Total power per core

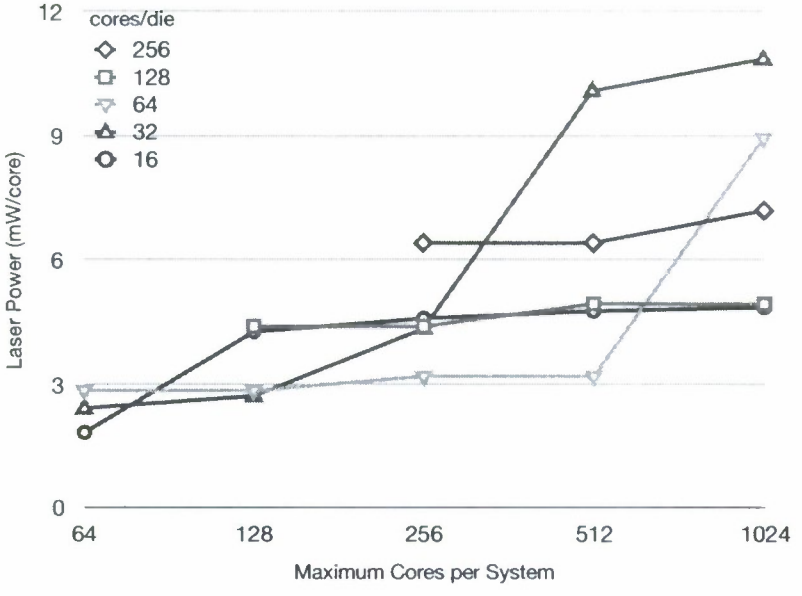


Figure 4.6: Laser power per core

Figure 4.6 shows the laser power required for fully populated systems as a function of die size and maximum system size. Systems that are not fully populated require the same laser power per core except for when only 1 or 2 sockets are populated and the star coupler is not needed. For all die sizes, as the maximum supported system size is increased, the required optical power is also increased, as expected. The rate at which it increases can fluctuate significantly because as the system size increases, some components (mixers, ADMs, star fiber couplers) are added to the interconnect and the loss rates of these components varies. A more interesting trend is that smaller systems are more efficiently constructed from smaller dies, as is visible on the pareto-optimal curve (underside of the graph). This appears to indicate that systems with a moderate number of sockets perform best because of the fan-out costs associated with making the all to all connectivity. With our selected technology, smaller dies have an advantage of shorter waveguides (less loss) as shown by the line for 16 cores per die.

4.3.2 Latency

Surprisingly latency does not get much worse when breaking sockets apart into smaller ones, even if electrical links are used off chip. As visible back in Table 2.2, both technologies get faster off chip after a minimum distance has been traversed to make up for the conversion delay. Once the overhead of getting onto a fiber is paid, the signal can travel 8cm in a clock cycle of our baseline system, so within less than a few cycles, everything is reachable by everything else on board. The only time link latency is worrisome is when trying to route a signal for a long distance electrically with a normal repeated wire on-chip, but this does not happen in our design since all long links are done photonically. Even with systems larger than the one in Section 3.1, the latency will not

get much bigger. With the largest conceivable board layouts, the link latency will still be less than 2ns, which will be dwarfed by the serialization latency.

4.3.3 Area

In general our photonic interconnect fits well within an area budget as shown in Figure 4.7. These are for die designs that are capable of supporting up to a maximum of 1024 cores in the system. Since our technology is using projected values, these overheads could change, but we are pessimistic in our assumptions about sizing, which results in over-estimates for area. Smaller dies use less area for the interconnect, because more of it is off chip and they are small enough that it is still possible to put many or all of the waveguides over the same air trenches. Although this suggests less wasted area is another reason smaller dies will be more cost-effective, the most important result is that using smaller dies is no worse than using larger ones, with respect to area overhead.

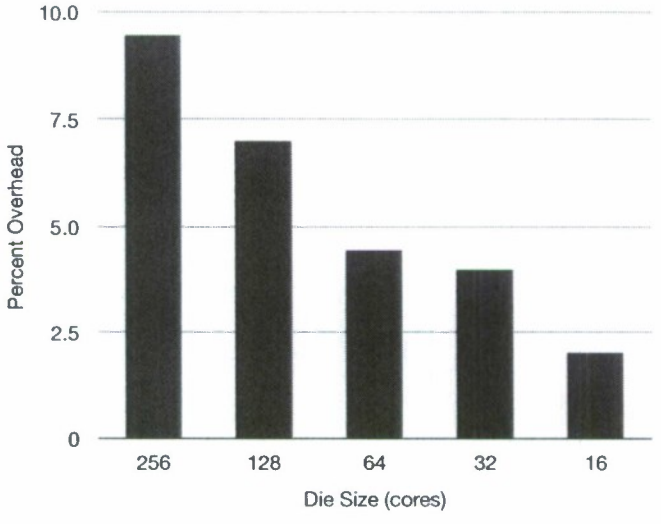


Figure 4.7: Percentage die area taken by photonic network (not including switch area)

4.3.4 Discussion

The main lesson is that there is a tradeoff between integration and disintegration. The models may not be able to fully capture all of the penalties of having many small dies, but dies smaller than the ones currently used may be feasible to make systems with a moderate number of sockets.

Historically systems have been built out of dies as large as is reasonable to manufacture because of the interconnect penalties of traversing socket boundaries. This sometimes results in paying a significant premium to fabricate larger monolithic dies. The photonic network template presented could reduce the barrier to multi-socket designs, enabling a new system design methodology of picking a die size that is cheapest to manufacture, and then using as many dies as needed to build the desired system.

Chapter 5

Related Work

The work of Batten et al. [4], identified the potential for monolithic silicon photonics for making an interconnection network to connect DRAM to processing cores. We used their technology assumptions and baseline machine as a starting point for our work. Our work differs in that it adds the contributions of using multi-socket systems as a way to reduce cost and considering coherence much more closely. Much of the other related work focusses on the on-chip network for a single chip and does not consider anything off-chip [7, 13, 16, 22, 24].

Kirman et al. presented a photonic on-chip interconnect in [16]. Their architecture attempted to utilize each interconnection topology for the range of distances it was best at. They subdivided a CMP into four blocks and those four blocks were connected by a photonic ring topology. Within a block electrical interconnects were used at a distance where they were advantageous to optical. Our network topologies were influenced by this, but we have made a more ambitious design that uses a more optimistic photonic technology.

Shacham et al. present a photonic NoC for a multiprocessor system that uses photonic switches built from crossings and resonant rings [24]. To set up a link, an electric control signal must travel ahead in parallel to the path to set up the switches. This enables them to get higher bandwidth utilization on their links than a point to point system like what was presented in this paper, but at the cost of path set up latency and the possibility of network contention. As a consequence of the set up requirements, they get the best performance from lightly contested bulk transfers.

Joshi et al. present a low-diameter photonic Clos network and compare it to electrical alternatives [13]. Their low-diameter network is motivated by the same desire of this work to provide uniform bandwidth while taking advantage of the distance insensitivity of photonic links. Unlike this work, their Clos network is able to utilize path diversity, but this would be harder to implement for the multisocket case because there are more endpoints to connect.

Phastlane [7] intends to bring the benefits of photonics to a dimension ordered mesh network. Since light propagates quickly, they allow a packet to sometimes travel multiple hops in a single cycle. Unlike [24], they set up each hop with an optical control signal that travels in parallel to the data payload. When there is contention, the packet will travel less hops in a cycle and is stored in an electrical buffer. If the buffer is full, the packet is dropped, and the sender is notified. Of the photonic proposals, Phastlane is the only one to consider not providing reliable transmission at the link layer.

Firefly [22] presents a hierarchical NoC. Similar to [16], it subdivides the chip into clusters, and within a cluster it uses electrical links and between clusters it uses photonic links. The photonic

links use a crossbar, but to prevent the need for global arbitration they break it up into multiple logical crossbars.

Proximity interconnect [9] is an interesting technology that is trying to solve many of the same problems our photonic socket-level interconnect is. It places dies very close together and uses capacitive coupling to transmit data without actual wire contacts. By doing so, it is able to obtain pitches and bandwidths comparable to on-chip wires. They have aspirations similar to ours for its use whether it be making small dies to reduce cost or combining large dies to approach wafer scale integration. Photonics, especially with DWDM should be able to achieve even higher bandwidths and is a little more robust of a technology since the exact relative alignment of two dies does not matter as much.

Three dimensional die stacking is another technology with the same motivation, but it could be used in conjunction with a photonic interconnect like in Corona [25]. They place their photonic network on its own die to give them more area and let them use better photonic materials which allows them to build more complicated networks. They use a large serpentine crossbar which has orders of magnitude more components than our networks and would be infeasible with our monolithically integrated photonics technology. As such, they burn significantly more laser power than our design for comparable bandwidth, but it is hard to accurately make this comparison since they are using a different photonic technology.

Chapter 6

Conclusion

In this work, we present design techniques that produce a general network template that can be scaled to handle varying numbers of cores and sockets. To scale our network design to even larger core counts will probably require moving to a multi-hop network.

Chip disintegration may seem counterintuitive for performance reasons, but with our photonic network, the performance degradation is made small enough that the cost incentives out weigh it. This could allow for a re-thinking of the design process where systems are built out of the appropriate number of the most economically sized die.

Due to the current state of silicon photonic research, multi-socket memory interconnects are a great application. In the near horizon, photonics provides great advantages over electrical at the scale of on-board/off-chip. To optimize these multi-socket systems, photonics should be used to communicate directly with DRAM, which will remove the last bit of wasteful off-chip electrical signaling. Further advances, such as efficient integrated lasers, will enable photonics research to continue to decrease the scale at which optical communication is advantageous, possibly opening up the chip micro-architecture as the next interesting application.

Appendix A

Coherence Considerations

To make this system more realizable it will need a coherency scheme (protocol and hardware implementation) to turn the network into a memory interconnect, which is something past designs have not given much consideration to. Especially for the general architecture presented in this paper, it is essential that the coherency scheme achieve the same goals of reusability and scalability. We want the same design to be able to handle different binary amounts of populated sockets in the system without unreasonable overheard. Our system uses shared memory, and coherency is maintained amongst all caches by a two level protocol corresponding to within and between clusters.

A.1 Intra-Cluster Coherence

Within a cluster, each core has its own private L1 cache and they all communicate through a shared L2 cache. The L2 cache is not inclusive of the L1s, but it does store duplicates of the tags. We envision using this with a protocol similar to what was described in Piranha [2]. This protocol will be responsible for keeping the caches within each cluster coherent, and requests that it cannot handle will be passed up to the next level coherency protocol.

A.2 Inter-Cluster Coherence

To maintain coherence between clusters we use a 4–hop MESI directory protocol. From the point of view of the directory, all caches in a cluster are lumped together and treated as one. We position a directory by every memory controller so it can intercept requests to memory and take the appropriate protocol actions. A directory is only responsible for the memory locations its associated memory controller provides. The protocol uses 4 hops because there is no core to core network, so all inter–core traffic must be routed through the memory controller.

To make the directory small enough to fit on chip rather than off-chip DRAM, we use a reverse tagged directory implemented with a Content Addressable Memory (CAM). For every cache line it is responsible for, the directory contains a duplicate of the cache tag and a few bits of protocol state. We reduce the associativity required for the directory by implementing it with many small CAMs where each one corresponds to a cache set. When a request is being looked up, only the CAM corresponding to the request's set needs to be examined. A cache tag's location in the reverse directory implicitly identifies the location of its owner. Because all the caches in the system are set associative, this puts a limit on the number of possible cache lines that could hold a block, namely

Nk if the system has N clusters and each one is k-way set associative. If this associativity is still too high, multiple CAM arrays could be used which will still be faster and cheaper than going to a direct mapped directory implemented by off-chip DRAM.

Although photonics provides great bandwidth which might tempt one to snoop, the energy cost at the endpoints to do associative lookups for every message at every cluster in the system will be prohibitive, especially as it scales. With snooping, for a given protocol miss (like a write miss), rather than searching the state of one cluster and the home directory, every cluster will need to be searched. This will also require a broadcast mechanism, which our current network topology does not provide. It could be possible to design it, but our topology was designed to minimally meet our goals and our coherency protocol works well without it. The bandwidth savings a directory protocol provides will also help the system scale to higher core counts and conserve energy.

A.3 Reusability

To support a variable number of populated sockets the way memory addresses are interleaved can be leveraged. For a given die size, if the number of populated sockets is doubled, the number of cache lines double, however the number of sets per cache that can address a particular memory controller get halved, so the number of possible locations a directory needs to be concerned with stays the same. The only thing that changes is the implicit addressing of clusters to tags in the reverse directory.

A.4 Directory Implementation Feasibility

To prove the feasibility of such a technique, we present a rough model of what reverse tagged directories would cost by scaling [5] down to 22 nm. To stress our design, we target the maximum size system our network targeted: 1024 cores over 1600 mm^2 of silicon. The target system uses a 48-bit physical address. Each cluster has 4MB of L2 cache that is 8-way set associative.

To implement the CAMs efficiently, we use a pre-computation based CAM [19] with a Half-NOR cell size of 0.34 μm^2 and a NAND cell size of 0.3695 μm^2. For the CAM arrays alone, this would take 50.531 mm^2, so rounding up generously for extra decode and control logic, this could be implemented in 80 mm^2 which is only 5% of the total silicon area.

The power required is harder to estimate due to its dependence on workload and coherence traffic. In 45 nm[5] each search took $0.14\frac{fJ}{bit}$, so including decode overheads and pessimistic energy scaling $0.1\frac{fJ}{bit}$ might be possible in 22 nm. Assuming the wildly high coherence activity rate of each core needing to access the directory once every five instructions results in 0.786W total. This amount will almost surely be drowned out by static power of the SRAMs included to hold the CAMs' state. The dynamic search power makes up such a small portion of the directory's power because the cache set partitioning makes the relative activity factor of any CAM cell quite low.

The latency of the directory itself should be quite tolerable. Even without much speed improvement from process technology and accounting for controller overhead, it should be possible to get a search done in under a nanosecond [5]. This should clearly win by more than an order of magnitude compared to off-chip DRAM. Overall we believe we could make an effective coherence mechanism utilizing reverse tagged directories built from on-chip CAMs.

Bibliography

[1] Krste Asanovic et al. The landscape of parallel computing research: A view from berkeley. Technical report, U.C. Berkeley, 2006.

[2] L. Barroso, K. Gharachorloo, R. McNamara, and A Nowatzyk et al. Piranha: A scalable architecture based on single-chip multiprocessing. *ISCA*, Jan 2000.

[3] T. Barwicz et al. Silicon photonics for compact, energy-efficient interconnects. *Journal of Optical Networking*, 6(1):63–73, 2007.

[4] C Batten, A Joshi, J Orcutt, A Khilo, B Moss, Charles Holzwarth, Miło s Popovic, Hanqing Li, Henry Smith, Judy Hoyt, Franz Kartner, Rajeev Ram, Vladimir Stojanovic, and Krste Asanovic. Building manycore processor-to-dram networks with monolithic silicon photonics. *High Performance Interconnects*, Jan 2008.

[5] Scott Beamer and Mehmet Akgul. Design of a low power content addressable memory (cam). *EE 241 Final Project*, May 2009.

[6] M. Brière, B. Girodias, Y. Bouchebaba, G. Nicolescu, F. Mieyeville, F. Gaffiot, and I. O'Connor. System level assessment of an optical noc in an mpsoc platform. In *DATE '07: Proceedings of the conference on Design, automation and test in Europe*, pages 1084–1089, San Jose, CA, USA, 2007. EDA Consortium.

[7] MJ Cianchetti, JC Kerekes, and DH Albonesi. Phastlane: a rapid transit optical routing network. *ISCA*, 36, 2009.

[8] William James Dally and Brian Towles. *Principles and Practices of Interconnection Networks*. Morgan Kaufmann, 1st edition, 2004.

[9] R Drost, R Hopkins, R Ho, and I Sutherland. Proximity communication. *IEEE Journal of Solid-State Circuits*, 39(9):1529 – 1535, Sep 2004.

[10] C. Holzwarth et al. Localized substrate removal technique enabling strong-confinement microphotonics in bulk si cmos processes. *Conf. on Lasers and Electro-Optics*, 2008.

[11] C. Gunn. CMOS photonics for high-speed interconnects. *IEEE Micro*, 26(2):58–66, Mar-Apr 2006.

[12] J. Hennessy and D. Patterson. *Computer Architecture: A Quantitative Approach*. Morgan Kaufmann, 4th edition, 2007.

[13] A Joshi, C Batten, Y Kwon, S Beamer, Imran Shamim, Krste Asanovic, and Vladimir Stojanovic. Silicon-photonic clos networks for global on-chip communication. *NOCS*, 3, Jan 2009.

[14] Ron Kalla. Power7: Ibm's next generation power microprocessor. *A Symposium on High Performance Chips*, 21, 2009.

[15] L. Kimerling et al. Electronic-photonic integrated circuits on the CMOS platform. *Proceedings of the SPIE*, 6125, Mar 2006.

[16] N Kirman, M Kirman, R Dokania, and J Martinez. Leveraging optical technology in future bus-based chip multiprocessors. *IEEE Micro*, 27(6), Jan 2006.

[17] Poonacha Kongetira, Kathirgamar Aingaran, and Kunle Olukotun. Niagara: A 32-way multithreaded sparc processor. *IEEE Micro*, page 9, Apr 2005.

[18] Sailesh Kottapalli and Jeff Baxter. Nhm-ex cpu architecture. *A Symposium on High Performance Chips*, 21, 2009.

[19] Chi-Sheng Lin, Jui-Chuan Chang, and Bin-Da Liu. A low-power precomputation-based fully parallel content-addressable memory. *JSSC*, 38(4):654– 662, 2003.

[20] M. Lipson. Compact electro-optic modulators on a silicon chip. *Journal of Sclected Topics in Quantum Electronics*, 12(6):1520–1526, Nov-Dec 2006.

[21] J. Orcutt et al. Demonstration of an electronic photonic integrated circuit in a commercial scaled bulk cmos process. *Conf. on Lasers and Elcctro-Optics*, 2008.

[22] Y Pan, P Kumar, J Kim, G Memik, Y Zhang, and A Choudhary. Firefly: illuminating future network-on-chip with nanophotonics. *ISCA*, 36, 2009.

[23] Sanjay Patel, Stephen Phillips, and Allan Strong. Sun's next-generation multi-threaded processor - rainbow falls. *A Symposium on High Performance Chips*, 21, 2009.

[24] A Shacham, B Lee, A Biberman, and K Bergman. Photonic noc for dma communications in chip multiprocessors. *IEEE Symposium High-Performance Interconnects*, 15, Jan 2007.

[25] D Vantrease, R Schreiber, M Monchiero, and M McLaren. Corona: System implications of emerging nanophotonic technology. *ISCA*, Jan 2008.

[26] M. Watts et al. Design, fabrication, and characterization of a free spectral range doubled ring-resonator filter. *Conf. on Lasers and Electro-Optics*, 1:269–272, May 2005.

[27] Samuel Webb Williams, Andrew Waterman, and David Patterson. Roofline: An insightful visual performance model for floating-point programs and multicore architectures. Technical Report UCB/EECS-2008-134, EECS Department, University of California, Berkeley, Oct 2008.

[28] Lei Zhang, Mei Yang, Yingtao Jiang, Emma Regentova, and Enyue Lu. Generalized wavelength routed optical micronetwork in network-on-chip. In *Proceedings of the 18th IASTED International Conference Parallel and Distributed Computing Systems*, 2006.